

# Optimal pattern matching algorithms

Gilles Didier

Aix-Marseille Université, CNRS, Centrale Marseille, I2M UMR7373, Marseille, France

E-mail: [gilles.didier@univ-amu.fr](mailto:gilles.didier@univ-amu.fr)

May 3, 2016

## Abstract

We study a class of finite state machines, called *w-matching machines*, which yield to simulate the behavior of pattern matching algorithms while searching for a pattern  $w$ . They can be used to compute the asymptotic speed, i.e. the limit of the expected ratio of the number of text accesses to the length of the text, of algorithms while parsing an iid text to find the pattern  $w$ .

Defining the order of a matching machine or of an algorithm as the maximum difference between the current and accessed positions during a search (standard algorithms are generally of order  $|w|$ ), we show that being given a pattern  $w$ , an order  $k$  and an iid model, there exists an optimal  $w$ -matching machine, i.e. with the greatest asymptotic speed under the model among all the machines of order  $k$ , of which the set of states belongs to a finite and enumerable set.

It shows that it is possible to determine: 1) the greatest asymptotic speed among a large class of algorithms, with regard to a pattern and an iid model, and 2) a  $w$ -matching machine, thus an algorithm, achieving this speed.

## 1 Introduction

The problem of pattern matching consists in reporting all, and only the occurrences of a (short) word, a *pattern*,  $w$  in a (long) word, a *text*,  $t$ . This question dates back to the early days of computer science. Since then, dozens of algorithms have been, and are still proposed, to solve it [6]. Pattern matching has a wide range of applications: text, signal and image processing, database searching, computer viruses detection, genetic sequences analysis etc. Moreover, as a classic algorithmic problem, it may serve to introduce new ideas and paradigms in this field. Though optimal algorithms, in the sense of worst case analysis, have been developed forty years ago [9], there exists as yet no algorithm which is fully efficient in all the various situations encountered in practice: large or small alphabets, long or short patterns etc. (see [6]).

The worst case analysis does not say much about the general behavior of algorithm in practical situations. In particular, the Knuth-Morris-Pratt algorithm is not much faster than the naive one, and even slower in average than certain algorithms with quadratic worst case complexity. A more accurate measure of the algorithm efficiency in real situations is the average complexity on random texts or, equivalently, the expected complexity under a probabilistic model of text. The question of average case analysis of pattern matching algorithms was raised since at least [9], in which the complexity of pattern matching algorithms is conveniently expressed in terms of number of text accesses. A seminal work shows that, under the assumption that both the symbols of the pattern  $w$  and the text are independently drawn uniformly from a finite alphabet, the minimum expectation of text accesses needed to search  $w$  in  $t$  is  $\mathbf{O}\left(\frac{|t| \log |w|}{|w|}\right)$  [18]. Since then, several works studied the average complexity of some pattern matching algorithms, mainly Boyer-Moore-Horspool and Knuth-Morris-Pratt [18, 8, 2, 1, 10, 15, 16, 17, 12, 13, 14, 11]. Different tools have been used to carry out these analysis, notably generating functions and Markov chains. In particular, G. Barth used Markov chains to compare the Knuth-Morris-Pratt and the naive algorithms [2]. More recently, T. Marschall and S. Rahmann provided a general framework based on the same underlying ideas, for performing statistical analysis of pattern matching algorithms, notably for computing the exact distributions of the number of text accesses of several pattern matching algorithms on iid texts [12, 13, 14, 11].

Following the same ideas, we consider finite state machines, called a *w-matching machines*, which yield to simulate the behavior of pattern matching algorithms while searching a given pattern  $w$ . They are used for studying the asymptotic behavior of pattern matching algorithms, namely the limit expectation of the ratio of the text length to the number of text accesses performed by an algorithm for searching a given pattern  $w$  in iid texts, which we call the *asymptotic speed* of the algorithm with regard to  $w$  and the iid model. We show that the sequence of states of a  $w$ -matching machine parsed while searching in an iid text follows a Markov chain, which yields to compute their asymptotic speed.

We next focus our interest in optimal  $w$ -matching machines, i.e. those with the greatest asymptotic speed with regard to an iid model (and the pattern  $w$ ). The order of a  $w$ -matching machine (or of an algorithm with regard to  $w$ ) is defined as the maximum difference between the current and the accessed positions during a search. Most of the  $w$ -matching machines corresponding to standard algorithms are of order  $|w|$  (a few of them have order  $|w| + 1$ ). We prove that, being given a pattern  $w$ , an order  $k$  and an iid model, there exists an optimal  $w$ -matching machine of order  $k$  in which the set of states is in bijection with the set of partial functions from  $\{0, \dots, k\}$  to the alphabet. It makes it possible to compute the greatest speed which can be achieved under a large class of algorithms (including all the pre-existing algorithms), and a  $w$ -machine achieving this speed. This optimal matching machine can be seen as a *de novo* pattern matching algorithm which is optimal with regard to the pattern and

the model. Some of the methods presented here have been implemented in the companion paper [5]. The software is available at <https://github.com/gilles-didier/Matchines.git>.

The rest of the paper is organized as follows. Section 2 gives some basic notations and definitions. In Section 3, we present the  $w$ -matching machines and some of their properties. Next, we present three standard probabilistic models of text, define the asymptotic speed of an algorithm and show that the sequence of internal states of a  $w$ -matching machine follows a Markov chain on iid texts (Section 4). Last, in Section 5, we show that any  $w$ -matching machine can be “simplified” into a no slower  $w$ -matching machine of order  $k$  with a set of states in bijection with the set of partial functions from  $\{0, \dots, k\}$  to the alphabet.

## 2 Definitions and notations

An *alphabet* is a finite set  $\mathcal{A}$  of elements called *letters* or *symbols*.

A *word*, a *text* or a *pattern* on  $\mathcal{A}$  is a finite sequence of symbols of  $\mathcal{A}$ . We put  $|v|$  for the length of a word  $v$  and  $|v|_w$  for the number of occurrences of the word  $w$  in  $v$ . The cardinal of the set  $\mathcal{S}$  is also noted  $|\mathcal{S}|$ . Words are indexed from 0, i.e.  $v = v_0v_1 \dots v_{|v|-1}$ . We put  $v_{[i,j]}$  for the subword of  $v$  starting at the position  $i$  and ending at the position  $j$ , i.e.  $v_{[i,j]} = v_iv_{i+1} \dots v_j$ . The concatenate of two words  $u$  and  $v$  is the word  $uv = u_0u_1 \dots u_{|u|-1}v_0v_1 \dots v_{|v|-1}$ .

For any length  $n \geq 0$ , we put  $\mathcal{A}^n$  for the set of words of length  $n$  on  $\mathcal{A}$  and  $\mathcal{A}^*$ , for the set of finite words on  $\mathcal{A}$ , i.e.  $\mathcal{A}^* = \bigcup_{n=0}^{\infty} \mathcal{A}^n$ .

Unless otherwise specified, all the texts and patterns considered below are on a fixed alphabet  $\mathcal{A}$ .

## 3 Matching machines

Let  $w$  be a pattern on an alphabet  $\mathcal{A}$ . A  $w$ -*matching machine* is 6-uple  $(Q, o, F, \alpha, \delta, \gamma)$  where

- $Q$  is a finite number of states,
- $o \in Q$  is the initial state,
- $F \subset Q$  is the subset of pre-match states,
- $\alpha : Q \rightarrow \mathbb{N}$  is the next-position-to-check function, which is such that for all  $q \in F$ ,  $\alpha(q) < |w|$ ,
- $\delta : Q \times \mathcal{A} \rightarrow Q$  is the transition state function,
- $\gamma : Q \times \mathcal{A} \rightarrow \mathbb{N}$  is the shift function.

By convention, the set of states of a matching machine always contains a *sink state*  $\odot$ , which is such that, for all symbols  $x \in \mathcal{A}$ ,  $\delta(\odot, x) = \odot$  and  $\gamma(\odot, x) = 0$ .

The *order*  $O_\Gamma$  of a matching machine  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  is its greatest next-position-to-check, i.e.  $O_\Gamma = \max_{q \in Q} \{\alpha(q)\}$ .

Remark that  $(Q, o, F, \delta)$  is a deterministic finite automaton. The  $w$ -matching machines carry the same information as the *Deterministic Arithmetic Automata* defined in [13, 14].

### 3.1 Generic algorithm

Algorithm 1, which will be referred to as the *generic algorithm*, takes a  $w$ -matching machine and a text  $t$  as input and is expected to output all the occurrence positions of  $w$  in  $t$ .

**input** : a  $w$ -matching machine  $(Q, o, F, \alpha, \delta, \gamma)$  and a text  $t$   
**output**: all the occurrence positions of  $w$  in  $t$  (hopefully)

```

 $(q, p) \leftarrow (o, 0)$ 
while  $p \leq |t| - |w|$  do
  | if  $q \in F$  and  $t_{p+\alpha(q)} = w_{\alpha(q)}$  then
  | | print “ occurrence at position  $p$  ”
  |  $(q, p) \leftarrow (\delta(q, t_{p+\alpha(q)}), p + \gamma(q, t_{p+\alpha(q)}))$ 

```

**Algorithm 1:** Generic algorithm

We put  $\mathbf{q}_\Gamma^t(i)$  (resp.  $\mathbf{p}_\Gamma^t(i)$ ) for the state  $q$  (resp. for the position  $p$ ) at the beginning of the  $i^{\text{th}}$  iteration of the generic algorithm on the input  $(\Gamma, t)$ . We put  $\mathbf{s}_\Gamma^t(i)$  for the shift at the end of the  $i^{\text{th}}$  iteration, i.e.  $\mathbf{s}_\Gamma^t(i) = \gamma(\mathbf{q}_\Gamma^t(i), t_{\mathbf{p}_\Gamma^t(i) + \alpha(\mathbf{q}_\Gamma^t(i))})$ . By convention, the generic algorithm starts with the iteration 0.

A  $w$ -matching machine  $\Gamma$  is *redundant* if there exist a text  $t$  and two indexes  $i < j$  such that

$$\mathbf{p}_\Gamma^t(j) + \alpha(\mathbf{q}_\Gamma^t(j)) = \mathbf{p}_\Gamma^t(i) + \alpha(\mathbf{q}_\Gamma^t(i)) + \sum_{k=i}^{j-1} \mathbf{s}_\Gamma^t(k).$$

In plain text, a matching machine  $\Gamma$  is redundant if there exists a text  $t$  for which a position is accessed more than once during an execution of the generic algorithm on the input  $(\Gamma, t)$ .

A  $w$ -matching machine  $\Gamma$  is *valid* if, for all texts  $t$ , the execution of the generic algorithm on the input  $(\Gamma, t)$  outputs all, and only the occurrence positions of  $w$  in  $t$ .

**Remark 1.** *If a matching machine is valid then*

1. *its order is greater than or equal to  $|w| - 1$ ,*

2. there is no text  $t$  such that for some  $j > i$ , we have  $\mathbf{q}_\Gamma^t(i) = \mathbf{q}_\Gamma^t(j)$  and  $\gamma(\mathbf{q}_\Gamma^t(k), t_{\mathbf{p}_\Gamma^t(k) + \alpha(\mathbf{q}_\Gamma^t(k))}) = 0$  for all  $i \leq k \leq j$ . In particular, the sink state is never reached during an execution of the generic algorithm with a valid machine.

*Proof.* Condition 1 comes from the fact that it is necessary to check the position  $(i + |w| - 1)$  of the text to make sure whether  $w$  occurs at  $i$  or not. If Condition 2 is not fulfilled, an infinite loop starts at the  $i^{\text{th}}$  iteration of the generic algorithm on the input  $(\Gamma, t)$ . In particular, the last occurrence of the pattern  $w$  in the text  $tw$  will never be reported.  $\square$

A *match transition* is a transition going from a state  $q \in F$  to the state  $\delta(q, w_{\alpha(q)})$ . It corresponds to an actual match if the machine is valid.

A  $w$ -matching machine  $\Gamma$  is *equivalent* to a  $w$ -matching machine  $\Gamma'$  if, for all texts  $t$ , the text accesses performed by the generic algorithm on the input  $(\Gamma, t)$  are the same as those performed on the input  $(\Gamma', t)$ . The machine  $\Gamma$  is *faster* than  $\Gamma'$  if, for all texts  $t$ , the number of iterations of the generic algorithm on the input  $(\Gamma, t)$  is smaller than that on the input  $(\Gamma', t)$ .

We claim that, for all pre-existing pattern matching algorithms and all patterns  $w$ , there exists a  $w$ -matching machine  $\Gamma$  which is such that the text accesses performed by the generic algorithm on the input  $(\Gamma, t)$  are the exact same as those performed by the pattern matching algorithm on the input  $(w, t)$ . Without giving a formal proof, this holds for any algorithm such that:

1. The current position in the text is stored in an internal variable which never decreases during their execution.
2. All the other internal variables, which will be refer to as *state variables*, are bounded independently of the texts in which the pattern is searched.
3. The difference between the position accessed and the current position only depends on the state variables.

We didn't find a pattern matching algorithm which not satisfies the conditions above.

Being given a pattern  $w$ , let us consider the  $w$ -matching machine where the set of states is made of the combinations of the possible values of the state variables, which are in finite number from Feature 2. Feature 3 ensures that we can define a next-position-to-check from the states of the machine, which is bounded independently from the input text. Last, the only changes which may occur between two text accesses during an execution of the algorithm, are an increment of the current position (Feature 1) and/or a certain number of modifications of the state variables, which ends up to change the state of the  $w$ -matching machine. For instance the  $w$ -matching machine  $\Gamma$  associated to the naive algorithm has  $|w|$  states with

- $Q = \{q_0, \dots, q_{|w|-1}\},$
- $o = q_0,$

- $F = \{q_{|w|-1}\},$
- $\alpha(q_i) = i$  for all indexes  $0 \leq i < w,$
- $\delta(q_i, a) = \begin{cases} q_{i+1} & \text{if } i < |w| - 1 \text{ and } a = w_i, \\ q_0 & \text{otherwise,} \end{cases}$
- $\gamma(q_i, a) = \begin{cases} 0 & \text{if } i < |w| - 1 \text{ and } a = w_i, \\ 1 & \text{otherwise.} \end{cases}$

A state  $q$  of the matching machine  $\Gamma$  is *reachable* in  $\Gamma$  if there exists a text  $t$  such that  $q$  is the current state of an iteration of the generic algorithm on the input  $(\Gamma, t)$ . Unless otherwise specified or for temporary constructions, we will only consider matching machines  $\Gamma$  in which all the states but the sink are reachable. Below, stating “removing all the unreachable states” will have to be understood as “removing all the unreachable states but the sink”. Remark that all reachable states  $q$  of a valid  $w$ -matching machine are such that there exists a text  $t$  and two indexes  $i \leq j$  such that  $\mathbf{q}_\Gamma^t(i) = q$  and  $\mathbf{q}_\Gamma^t(j) \in F$ . In the same way, a transition between two given states is reachable if there exists a text  $t$  for which the transition occurs during the execution of the generic algorithm on the input  $(\Gamma, t)$ .

We assume that for all pre-match states  $q$  of  $\Gamma$ , there exists a text  $t$  such that a match transition starting from  $q$  occurs during the execution of the generic algorithm on the input  $(\Gamma, t)$ .

### 3.2 Full-memory expansion – standard matching machines

For all positive integers  $n$ ,  $R_n$  denotes the set of subsets  $H$  of  $\{0, \dots, n\} \times \mathcal{A}$  such that, for all  $i \in \{0, \dots, n\}$ , there exists at most one pair in  $H$  with  $i$  as first entry. In other words,  $R_n$  is the set of partial functions from  $\{0, \dots, n\}$  to  $\mathcal{A}$ .

For  $H \in R_n$ , we put  $\mathbf{f}(H)$  for the set consisting of the first entries (i.e. the *position entries*) of the pairs in  $H$ , namely

$$\mathbf{f}(H) = \{i \mid \exists x \in \mathcal{A} \text{ with } (i, x) \in H\}.$$

Let  $k$  be a non-negative integer and  $H \in R_n$ , the  $k$ -shifted of  $H$  is defined by

$$\overset{k}{\leftarrow} H = \{(u - k, y) \mid (u, y) \in H \text{ with } u \geq k\}.$$

In plain text,  $\overset{k}{\leftarrow} H$  is obtained by subtracting  $k$  from the position entries of the pairs in  $H$  and by keeping only the pairs with non-negative positive entries.

The *full memory expansion* of a  $w$ -matching machine  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  is the  $w$ -matching machine  $\hat{\Gamma}$  obtained by removing the unreachable states of the  $w$ -matching machine  $\Gamma' = (Q', o', F', \alpha', \delta', \gamma')$ , defined as:

- $Q' = Q \times R_{O_\Gamma},$

- $o' = (o, \emptyset)$ ,
- $\alpha'((q, H)) = \alpha(q)$ ,
- $\gamma'((q, H), x) = \gamma(q, x)$ ,
- $F' = F \times R_{O_\Gamma}$ ,
- 

$$\delta'((q, H), x) = \begin{cases} (\delta(q, x), \overleftarrow{H \cup \{(\alpha(q), x)\}}^{\gamma(q, x)}) & \text{if } \alpha(q) \notin \mathbf{f}(H), \\ \odot & \text{if } \exists a \neq x \text{ s.t. } (\alpha(q), a) \in H, \\ (\delta(q, x), \overleftarrow{H}^{\gamma(q, x)}) & \text{if } (\alpha(q), x) \in H. \end{cases}$$

**Remark 2.** At the beginning of the  $i^{\text{th}}$  iteration of the generic algorithm on the input  $(\hat{\Gamma}, t)$ , if the current state is  $(q, H)$  then the positions of  $\{(j + \mathbf{p}_\Gamma^t(i)) \mid j \in \mathbf{f}(H)\}$  are exactly the positions of  $t$  greater than  $\mathbf{p}_\Gamma^t(i)$  which were accessed so far, while the second entries of the corresponding elements of  $H$  give the symbols read.

**Proposition 1.** The  $w$ -matching machines  $\Gamma$  and  $\hat{\Gamma}$  are equivalent. In particular  $\Gamma$  is valid (resp. non-redundant) if and only if  $\hat{\Gamma}$  is valid (resp. non-redundant).

*Proof.* It is straightforward to prove by induction that, for all iterations  $i$ , if  $\mathbf{q}_\Gamma^t(i) = (q, H)$  then  $\mathbf{q}_{\hat{\Gamma}}^t(i) = q$ , reciprocally, there exists  $H \in R_{O_\Gamma}$  such that  $\mathbf{q}_{\hat{\Gamma}}^t(i) = (\mathbf{q}_\Gamma^t(i), H)$ ,  $\mathbf{s}_\Gamma^t(i) = \mathbf{s}_{\hat{\Gamma}}^t(i)$  and  $\mathbf{p}_\Gamma^t(i) = \mathbf{p}_{\hat{\Gamma}}^t(i)$ .  $\square$

A  $w$ -matching machine  $\Gamma$  is *standard* if its set of states has the same cardinal as that of its full memory expansion or, equivalently, if each state  $q$  of  $\Gamma$  appears in a unique pair/state of its full memory expansion. For all states  $q$  of a standard matching machine  $\Gamma$ , we put  $\mathbf{h}_\Gamma(q)$  for the second entry of the unique pair/state of  $\hat{\Gamma}$  in which  $q$  appears.

**Remark 3.** Let  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  be a standard matching machine. For all paths  $q_0, \dots, q_n$  of states in  $Q_{\setminus \{\odot\}}$  in the DFA  $(Q, o, Q, \delta)$ , there exists a text  $t$  such that  $\mathbf{q}_\Gamma^t(i) = q_i$  for all  $0 \leq i \leq n$ .

**Theorem 1.** A standard and non-redundant  $w$ -matching machine  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  is valid if and only if, for all  $q \in Q$ ,

1.  $q \in F$  if and only if we have  $(j, w_j) \in \mathbf{h}_\Gamma(q)$  for all  $j \in \{0, \dots, |w| - 1\} \setminus \{\alpha(q)\}$ ,

2.

$$\begin{aligned} & \gamma(q, x) \\ & \leq \begin{cases} \min\{k \geq 1 \mid w_{i-k} = y \text{ for all } (i, y) \in \mathbf{h}_\Gamma(q) \cup \{(\alpha(q), x)\} \text{ with } k \leq i < k + |w|\} & \text{if } q \in F, \\ \min\{k \geq 0 \mid w_{i-k} = y \text{ for all } (i, y) \in \mathbf{h}_\Gamma(q) \cup \{(\alpha(q), x)\} \text{ with } k \leq i < k + |w|\} & \text{otherwise,} \end{cases} \end{aligned}$$

3. there is no path  $(q_0, \dots, q_\ell)$  such that

- $q_i \neq q_j$  for all  $0 \leq i \neq j \leq \ell$ ,
- there exists a word  $v$  such that
  - $q_{i+1} = \delta(q_i, v_i)$  and  $\gamma(q_i, v_i) = 0$  for all  $0 \leq i < \ell$ ,
  - $q_0 = \delta(q_\ell, v_\ell)$  and  $\gamma(q_\ell, v_\ell) = 0$ .

*Proof.* We recall our implicit assumption that all the states of  $Q$  are reachable. Let us assume that the property 1 of the theorem is not granted. Either there exists a state  $q \in F$  and a position  $j \in \{0, \dots, |w| - 1\} \setminus \{\alpha(q)\}$  such that  $(j, w_j) \notin \mathbf{h}_\Gamma(q)$  or there exists a state  $q \notin F$  with  $(j, w_j) \in \mathbf{h}_\Gamma(q)$  for all  $j \in \{0, \dots, |w| - 1\} \setminus \{\alpha(q)\}$ . From the implicit assumption, there exists a text  $t$  and an iteration  $i$  such that  $\mathbf{q}_\Gamma^t(i) = q$ . Since  $\Gamma$  is non-redundant, the position  $\mathbf{p}_\Gamma^t(i) + \alpha(q)$  was not accessed before the iteration  $i$  and we can assume that  $t_{\mathbf{p}_\Gamma^t(i) + \alpha(q)} = w_{\alpha(q)}$ . If  $q \in F$  then the generic algorithm reports an occurrence of  $w$  at  $\mathbf{p}_\Gamma^t(i)$ . Furthermore, since  $\Gamma$  is standard, if there exists  $j \in \{0, \dots, |w| - 1\} \setminus \{\alpha(q)\}$  such that  $(j, w_j) \notin \mathbf{h}_\Gamma(q)$ , then either the position  $\mathbf{p}_\Gamma^t(i) + j$  was accessed with  $t_{\mathbf{p}_\Gamma^t(i) + j} \neq w_j$  or it was not accessed and we can choose  $t_{\mathbf{p}_\Gamma^t(i) + j} \neq w_j$ . In both cases,  $w$  does not occur at  $\mathbf{p}_\Gamma^t(i)$  thus  $\Gamma$  is not valid. Let now assume that  $q \notin F$  and  $(j, w_j) \in \mathbf{h}_\Gamma(q)$  for all  $j \in \{0, \dots, |w| - 1\} \setminus \{\alpha(q)\}$ . This implies that  $w$  does occur at the position  $\mathbf{p}_\Gamma^t(i)$  which is not reported at the iteration  $i$ . Since, from the definition of  $w$ -matching machines, the states  $q'$  of  $F$  are such that  $\alpha(q') < |w|$  and  $\Gamma$  is non-redundant, the states parsed at iterations  $j > i$  and such that  $\mathbf{p}_\Gamma^t(j) = \mathbf{p}_\Gamma^t(i)$  are not in  $F$ . It follows that the position  $\mathbf{p}_\Gamma^t(i)$  is not reported at any further iteration. Again,  $\Gamma$  is not valid.

If the property 2 is not granted, it is straightforward to build a text  $t$  for which an occurrence position of  $w$  is not reported.

From the second item of Remark 1, if the property 3 is not granted then  $\Gamma$  is not valid.

Reciprocally, if  $\Gamma$  is not valid, there exist a text  $t$  and a position  $m$  for whose one of the following assertions holds:

1. the pattern  $w$  occurs at the position  $m$  of  $t$  and  $m$  is not reported by the generic algorithm on the input  $(\Gamma, t)$ ,
2. the generic algorithm reports the position  $m$  on the input  $(\Gamma, t)$  but the pattern  $w$  does not occurs at  $m$ .

Let us first assume that the generic algorithm is such that  $\mathbf{p}_\Gamma^t(i) < m$  for all iterations  $i$ . Considering an iteration  $i > (m + 1)(|Q| + 1)$ , there exists an iteration  $k \leq i$  such that for all  $0 \leq \ell \leq |Q| + 1$ , we have  $\mathbf{s}_\Gamma^t(k + \ell) = 0$ , i.e. there exists a path  $(q_0, \dots, q_\ell)$  negating the property 3.

Let us now assume that there exists an iteration  $i$  with  $\mathbf{p}_\Gamma^t(i) > m$  and let  $k$  be the greatest index such that  $\mathbf{p}_\Gamma^t(k) \leq m$ . If  $w$  occurs at the position  $m$  which is not reported then the fact that  $\mathbf{p}_\Gamma^t(k) \leq m$  and  $\mathbf{p}_\Gamma^t(k + 1) > m$  contradicts the property 2. Let us assume that  $w$  does not occur at  $m$  which is reported during the execution. We have necessarily  $\mathbf{p}_\Gamma^t(k) = m$ ,  $\mathbf{q}_\Gamma^t(k) \in F$  and  $\alpha(\mathbf{q}_\Gamma^t(k)) < |w|$ .



If  $w$  does not occur at  $m$ , then there exists a position  $j \in \{0, \dots, |w|-1\} \setminus \{\alpha(q)\}$  such that  $t_{\mathbf{p}_\Gamma^t(k)+j} \neq w_j$  and, from Remark 2, we get  $(j, w_j) \notin \mathbf{h}_\Gamma(\mathbf{q}_\Gamma^t(k))$  thus a contradiction with the property 1.  $\square$

Let  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  be a matching machine and  $\dot{q}$  and  $\ddot{q}$  be two states of  $Q$ . The *redirected matching machine*  $\Gamma_{\ddot{q} \triangleright \dot{q}}$  is constructed from  $\Gamma$  by redirecting all the transitions that end with  $\ddot{q}$ , to  $\dot{q}$ . Namely, the matching machine  $\Gamma_{\ddot{q} \triangleright \dot{q}}$  is obtained by removing the unreachable states of  $\Gamma' = (Q', o', F', \alpha', \delta', \gamma')$ , defined for all  $q \in Q \setminus \{\ddot{q}\}$  and all symbols  $x$ , as:

- $Q' = Q \setminus \{\ddot{q}\}$ ,
- $o' = \begin{cases} \dot{q} & \text{if } \ddot{q} = o, \\ o & \text{otherwise,} \end{cases}$
- $F' = \begin{cases} F \setminus \{\ddot{q}\} \cup \{\dot{q}\} & \text{if } \ddot{q} \in F, \\ F & \text{otherwise,} \end{cases}$
- $\alpha'(q) = \alpha(q)$ ,
- $\gamma' = \gamma(q, x)$ ,
- $\delta'(q, x) = \begin{cases} \dot{q} & \text{if } \delta(q, x) = \ddot{q}, \\ \delta(q, x) & \text{otherwise.} \end{cases}$

**Lemma 1.** *Let  $\Gamma$  be a standard  $w$ -matching machine and  $\dot{q}$  and  $\ddot{q}$  be two states of  $Q$  such that  $\mathbf{h}_\Gamma(\dot{q}) = \mathbf{h}_\Gamma(\ddot{q})$ . The redirected machines  $\Gamma_{\ddot{q} \triangleright \dot{q}}$  and  $\Gamma_{\dot{q} \triangleright \ddot{q}}$  are both standard. Moreover, if  $\Gamma$  is valid then both  $\Gamma_{\ddot{q} \triangleright \dot{q}}$  and  $\Gamma_{\dot{q} \triangleright \ddot{q}}$  are valid.*

*Proof.* The fact that  $\Gamma_{\ddot{q} \triangleright \dot{q}}$  and  $\Gamma_{\dot{q} \triangleright \ddot{q}}$  are standard comes straightforwardly from the fact that  $\mathbf{h}_\Gamma(\dot{q}) = \mathbf{h}_\Gamma(\ddot{q})$ .

Let us assume that  $\Gamma_{\ddot{q} \triangleright \dot{q}}$  is not valid. There exist a text  $t$  and a position  $m$  for whose one of the following assertions holds:

1. the pattern  $w$  occurs at the position  $m$  of  $t$  and  $m$  is not reported by the generic algorithm on the input  $(\Gamma_{\ddot{q} \triangleright \dot{q}}, t)$ ,
2. the generic algorithm reports the position  $m$  on the input  $(\Gamma_{\ddot{q} \triangleright \dot{q}}, t)$  but the pattern  $w$  does not occurs at  $m$ .

By construction, the smallest index  $k$  such that  $\mathbf{q}_{\Gamma_{\ddot{q} \triangleright \dot{q}}}^t(k) \neq \mathbf{q}_\Gamma^t(k)$  verifies  $\mathbf{q}_{\Gamma_{\ddot{q} \triangleright \dot{q}}}^t(k) = \dot{q}$ ,  $\mathbf{q}_\Gamma^t(k) = \ddot{q}$  and  $\mathbf{p}_{\Gamma_{\ddot{q} \triangleright \dot{q}}}^t(i) = \mathbf{p}_\Gamma^t(i)$  for all  $i \leq k$ . If there is no iteration  $j$  such that both  $\mathbf{q}_{\Gamma_{\ddot{q} \triangleright \dot{q}}}^t(j) = \dot{q}$  and  $\mathbf{p}_{\Gamma_{\ddot{q} \triangleright \dot{q}}}^t(j) \leq m$  then the executions of the standard algorithm coincide beyond the position  $m$  on the inputs  $(\Gamma, t)$  and  $(\Gamma_{\ddot{q} \triangleright \dot{q}}, t)$ . If  $\Gamma_{\ddot{q} \triangleright \dot{q}}$  is not valid then  $\Gamma$  is not valid.

Let us now assume that  $q$  is reached before parsing the position  $m$  on the input  $(\Gamma_{\ddot{q} \triangleright \dot{q}}, t)$  and let  $j$  be the greatest index such that  $\mathbf{q}_{\Gamma_{\ddot{q} \triangleright \dot{q}}}^t(j) = \dot{q}$  and  $\mathbf{p}_{\Gamma_{\ddot{q} \triangleright \dot{q}}}^t(j) \leq m$ . Since the state  $\dot{q}$  is reachable with  $\Gamma$ , there exists a text  $u$  and an index  $i$  such that  $\dot{q}$  is the current state of the  $i^{\text{th}}$  iteration of the standard

algorithm on the input  $(\Gamma, u)$ . Let now define  $v = u_{[0, \mathbf{p}_\Gamma^t(i)-1]} t_{[\mathbf{p}_{\Gamma_{\dot{q} \triangleright \dot{q}}}^t(j), |t|-1]}$ . Since  $\Gamma$  is standard, the positions greater than  $\mathbf{p}_\Gamma^u(i)$  accessed by the generic algorithm on the input  $(\Gamma, u)$  at the  $i^{\text{th}}$  iteration are  $\{(k + \mathbf{p}_\Gamma^t(i)) \mid k \in \mathbf{f}(\mathbf{h}_\Gamma(\dot{q}))\}$  and the positions greater than  $\mathbf{p}_{\Gamma_{\dot{q} \triangleright \dot{q}}}^t(j)$  accessed by the generic algorithm at the  $j^{\text{th}}$  on the input  $(\Gamma_{\dot{q} \triangleright \dot{q}}, t)$  iteration are  $\{(k + \mathbf{p}_{\Gamma_{\dot{q} \triangleright \dot{q}}}^t(j)) \mid k \in \mathbf{f}(\mathbf{h}_\Gamma(\dot{q}))\}$  (Remark 2). When considered relatively to the current positions  $\mathbf{p}_\Gamma^t(i)$  and  $\mathbf{p}_{\Gamma_{\dot{q} \triangleright \dot{q}}}^t(j)$ , the accessed positions greater than these current positions are the same. The positions accessed until the  $i^{\text{th}}$  iteration on the inputs  $(\Gamma, u)$  and  $(\Gamma, v)$ , coincide. In particular, we have  $\mathbf{q}_\Gamma^v(i) = \mathbf{q}_\Gamma^u(i) = \dot{q}$ . With the definitions of the text  $v$ , the execution of the generic algorithm from the  $i^{\text{th}}$  on the input  $(\Gamma, v)$  does coincide with the execution of from the  $j^{\text{th}}$  iteration on the input  $(\Gamma_{\dot{q} \triangleright \dot{q}}, t)$ . Again, if  $\Gamma_{\dot{q} \triangleright \dot{q}}$  is not valid then  $\Gamma$  is not valid.  $\square$

**Lemma 2.** *Let  $\Gamma$  be a  $w$ -matching machine which is both valid and standard. For all states  $q$  and all symbols  $x$  and  $y$ , if  $\delta(q, x) = \delta(q, y) \neq \odot$  then  $\gamma(q, x) = \gamma(q, y)$ .*

*Proof.* Let us first remark that, since  $\Gamma$  is standard, the fact that  $\delta(q, x) = \delta(q, y) \neq \odot$  implies that  $\mathbf{h}_\Gamma(\delta(q, x)) = \mathbf{h}_\Gamma(\delta(q, y))$ . It follows that both  $\gamma(q, x)$  and  $\gamma(q, y)$  are strictly greater than  $\alpha(q)$ .

Let  $d$  be the greatest position entry of the elements of  $\mathbf{h}_\Gamma(q) \cup \{(\alpha(q), x)\}$ , i.e.  $d = \max(\mathbf{f}(\mathbf{h}_\Gamma(q)) \cup \{\alpha(q)\})$ . By construction if  $\mathbf{h}_\Gamma(\delta(q, x))$  (resp.  $\mathbf{h}_\Gamma(\delta(q, y))$ ) is not empty then the greatest position entry of its elements is  $d - \gamma(q, x)$  (resp.  $d - \gamma(q, y)$ ). It follows that  $\mathbf{h}_\Gamma(\delta(q, x)) = \mathbf{h}_\Gamma(\delta(q, y))$  and  $\gamma(q, x) \neq \gamma(q, y)$  is only possible if  $\mathbf{h}_\Gamma(\delta(q, x)) = \mathbf{h}_\Gamma(\delta(q, y)) = \emptyset$ , which implies that both  $\gamma(q, x)$  and  $\gamma(q, y)$  are strictly greater than  $d$ .

Let us assume that  $\gamma(q, x) < \gamma(q, y)$ . We then have that  $\gamma(q, y) > d + 1$ . Let  $t$  be a text such that there is a position  $i$  with  $\mathbf{q}_\Gamma^t(i) = q$ ,  $t_{\mathbf{p}_\Gamma^t(i) + \alpha(q)} = y$  and  $w$  occurs at the position  $\mathbf{p}_\Gamma^t(i) + d$ . Such a text  $t$  exists since the state  $q$  is reachable (with our implicit assumption) and the only positions of  $t$  that we set, are not accessed until iteration  $i$ . Since  $\gamma(q, y) > d + 1$ , the occurrence of  $w$  at the position  $\mathbf{p}_\Gamma^t(i) + d$  cannot be reported, which contradicts the assumption that  $\Gamma$  is valid.  $\square$

### 3.3 Compact matching machines

A  $w$ -matching machine  $\Gamma$  is *compact* if it does not contain a state  $\dot{q}$  such that one of the following assertions holds:

1. there exists a symbol  $x$  with  $\delta(\dot{q}, x) \neq \odot$  and  $\delta(\dot{q}, y) = \odot$  for all symbols  $y \neq x$ ;
2. for all symbols  $x$  and  $y$ , we have both  $\delta(\dot{q}, x) = \delta(\dot{q}, y)$  and  $\gamma(\dot{q}, x) = \gamma(\dot{q}, y)$ .

Let  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  be a non-compact  $w$ -matching machine,  $\dot{q}$  be a state verifying one of the two assertions making  $\Gamma$  non-compact. If  $\dot{q}$  verifies

the assertion 1 and  $x$  is the only symbol such that  $\delta(\dot{q}, x) \neq \odot$ , we set  $\delta(\dot{q}, \cdot) = \delta(\dot{q}, x)$  and  $\gamma(\dot{q}, \cdot) = \gamma(\dot{q}, x)$ . If  $\dot{q}$  verifies the assertion 2, we set  $\delta(\dot{q}, \cdot) = \delta(\dot{q}, x)$  and  $\gamma(\dot{q}, \cdot) = \gamma(\dot{q}, x)$ , by picking any symbol  $x$ . The  $w$ -matching machine  $\Gamma_{\dot{q}} = (Q_{\dot{q}}, o_{\dot{q}}, F_{\dot{q}}, \alpha_{\dot{q}}, \delta_{\dot{q}}, \gamma_{\dot{q}})$  is defined, for all states  $q \in Q \setminus \{\dot{q}\}$  and all symbols  $x$ , as

- $Q_{\dot{q}} = Q \setminus \{\dot{q}\}$ ,
- $o_{\dot{q}} = \begin{cases} o & \text{if } \dot{q} \neq o, \\ \delta(\dot{q}, \cdot) & \text{otherwise,} \end{cases}$
- $F_{\dot{q}} = \begin{cases} F & \text{if } \dot{q} \notin F, \\ F \setminus \{\dot{q}\} \cup \{q \mid \exists x \in \mathcal{A} \text{ with } \delta(q, x) = \dot{q}\} & \text{otherwise,} \end{cases}$
- $\alpha_{\dot{q}}(q) = \alpha(q)$ ,
- $\delta_{\dot{q}}(q, x) = \begin{cases} \delta(q, x) & \text{if } \delta(q, x) \neq \dot{q}, \\ \delta(\dot{q}, \cdot) & \text{otherwise,} \end{cases}$
- $\gamma_{\dot{q}}(q, x) = \begin{cases} \gamma(q, x) & \text{if } \delta(q, x) \neq \dot{q}, \\ \gamma(q, x) + \gamma(\dot{q}, \cdot) & \text{otherwise.} \end{cases}$

If all the states of  $\Gamma$  are reachable, then so are all the states of  $\Gamma_{\dot{q}}$ .

The following lemma ensures that any standard machine can be made compact and that this operation cannot deteriorate its efficiency.

**Lemma 3.** *Let  $\Gamma$  be a  $w$ -matching machine which is made non-compact by a state  $\dot{q}$ .*

1. *If  $\Gamma$  is standard then  $\Gamma_{\dot{q}}$  is standard.*
2. *If  $\Gamma$  is valid then  $\Gamma_{\dot{q}}$  is valid.*
3.  *$\Gamma_{\dot{q}}$  is faster than  $\Gamma$ .*

*Proof.* We start by noting that if  $\Gamma$  is both standard and non-compact then there exist a state  $\dot{q}$  and a symbol  $x$  such that  $\delta(\dot{q}, x) \neq \odot$  and  $\delta(\dot{q}, y) = \odot$  for all symbols  $y \neq x$  (the other property leading to the non-compactness is excluded if  $\Gamma$  is standard). It follows that we have  $(\alpha(\dot{q}), x) \in \mathbf{h}_{\Gamma}(\dot{q})$ . Redirecting all the transitions that end with  $\dot{q}$ , to  $\delta(\dot{q}, x)$  and incrementing the shifts accordingly does not change the set  $\mathbf{h}_{\Gamma}(\delta(\dot{q}, x))$ , nor any set  $\mathbf{h}_{\Gamma}(q)$ . The matching machine  $\Gamma_{\dot{q}}$  is still standard.

Let now assume that  $\Gamma$  is valid. In particular, the transitions to the sink state are never encountered (Remark 1). By construction, the sequence of states parsed during an execution of the generic algorithm with  $\Gamma_{\dot{q}}$ , can be obtained by withdrawing all the positions in which  $\dot{q}$  occurs from the sequence observed with  $\Gamma$ . The machine  $\Gamma_{\dot{q}}$  is thus valid and faster than the initial one.  $\square$

**Remark 4.** *If a  $w$ -matching machine  $\Gamma$  is both standard and compact then it is not redundant.*

**Proposition 2.** *If  $\Gamma$  is a valid  $w$ -matching machine then there exists a standard, compact and valid  $w$ -matching machine  $\Gamma'$  which is faster or equivalent to  $\Gamma$ .*

*Proof.* By construction and from Proposition 1, the full memory expansion of  $\Gamma$  is both standard, valid and equivalent to  $\Gamma$ . Next, applying Lemma 3 as long as there exist a state  $q$  and a symbol  $x$  such that  $\delta(q, x) \neq \odot$  and  $\delta(q, y) = \odot$  for all symbols  $y \neq x$ , leads to a compact, standard and valid  $w$ -matching machine faster or equivalent to  $\Gamma$ .  $\square$

## 4 Random text models and asymptotic speed

### 4.1 Text models

A text model on an alphabet  $\mathcal{A}$  defines a probability distribution on  $\mathcal{A}^n$  for all lengths  $n$ . Two text models are said *equivalent* if they define the same probability distributions on  $\mathcal{A}^n$  for all lengths  $n$ .

We present three embedded classes of random text models, namely independent identically distributed, a.k.a. Bernoulli, Markov and Hidden Markov models.

An *independent identically distributed (iid) model* is fully specified by a probability distribution  $\pi$  on the symbols of the alphabet. It will be simply referred to as “ $\pi$ ”. Under the model  $\pi$ , the probability of a text  $t$  is

$$p_\pi(t) = \prod_{i=0}^{|t|-1} \pi(t_i).$$

A *Markov model*  $M$  of order  $n$  is a 2-uple  $(\pi_M, \delta_M)$ , where  $\pi_M$  is a probability distribution on the words of length  $n$  of the alphabet (the initial distribution) and  $\delta_M$  associates a pair made of a word  $u$  of length  $n$  and a symbol  $x$  with the probability for  $u$  to be followed by  $x$  (the transition probability). Under a Markov model  $M = (\pi_M, \delta_M)$  of order  $n$ , the probability of a text  $t$  of length greater than  $n$  is

$$p_M(t) = \pi_M(t_{[0, n-1]}) \prod_{i=n}^{|t|-1} \delta_M(t_{[i-n, i-1]}, t_i).$$

The probability distributions of words of length smaller than  $n$  are obtained by marginalizing the distribution  $\pi_M$ . Under this definition, Markov models are homogeneous (i.e. such that the transition probabilities do not depend on the position). “Markov model” with no order specified stands for “Markov model of order 1”.

A *Hidden Markov model (HMM)*  $H$  is a 4-uple  $(Q_H, \pi_H, \delta_H, \phi_H)$  where  $Q_H$  is a set of (hidden) states,  $(\pi_H, \delta_H)$  is a Markov model of order 1 on  $Q_H$ , and  $\phi_H$  associates a pair made of a state  $q$  and of a symbol  $x$  of the text alphabet with the probability for the state  $q$  to emit  $x$  (i.e.  $\phi_H(q, \cdot)$  is a probability

distribution on the text alphabet). Under a HMM  $H$ , the probability of a text  $t$  is

$$p_H(t) = \sum_{q \in Q_H^{|t|}} \pi_H(q_0) \phi_H(q_0, t_0) \prod_{i=1}^{|t|-1} \delta_H(q_{i-1}, q_i) \phi_H(q_i, t_i).$$

We will often consider HMMs  $H = (Q_H, \pi_H, \delta_H, \phi_H)$  with *deterministic emission functions*, i.e. such that for all states  $d \in Q_H$  there exists a unique symbol  $x$  with  $\phi_H(d, x) > 0$ , i.e. with  $\phi_H(d, x) = 1$ . In this case, for all states  $d$ , we will put  $\psi_H(d)$  for the unique symbol such that  $\phi_H(d, \psi_H(d)) > 0$  ( $\psi_H$  is just a map from  $Q_H$  to the alphabet). Remark that for all HMM  $H$ , there exists a HMM  $H'$  with a deterministic emission function which is equivalent to  $H$  (it is obtained by splitting the hidden states according to the symbols emitted and by setting the probability transitions accordingly). In [13, 14], authors define the *finite-memory text models* which are essentially HMMs with an additional emission function.

Basically, iid models are special cases of Markov models which are themselves special cases of HMMs.

The next theorem is essentially a restatement of Item 1 of Lemma 3 in [14], for matching machines and HMMs.

**Theorem 2** ([14]). *Let  $\Gamma = (Q, o, F, \delta, \alpha, \gamma)$  be a  $w$ -matching machine. If a text  $t$  follows an HMM then there exists a Markov model  $(\pi_{H'}, \delta_{H'})$  of state set  $Q_{H'}$  such that there exist:*

- a map  $\psi_{H'}^{[t]}$  from  $Q_{H'}$  to  $\mathcal{A}$  such that  $t$  follows the HMM with deterministic emission  $(Q_{H'}, \pi_{H'}, \delta_{H'}, \psi_{H'}^{[t]})$ ,
- a map  $\psi_{H'}^{[Q]}$  from  $Q_{H'}$  to  $Q$  such that  $(\mathbf{q}_\Gamma^t(i))_i$  follows the HMM with deterministic emission  $(Q_{H'}, \pi_{H'}, \delta_{H'}, \psi_{H'}^{[Q]})$
- a map  $\psi_{H'}^{[s]}$  from  $Q_{H'}$  to  $\{0, \dots, |w|\}$  such that  $(\mathbf{s}_\Gamma^t(i))_i$  follows the HMM with deterministic emission  $(Q_{H'}, \pi_{H'}, \delta_{H'}, \psi_{H'}^{[s]})$

*Proof.* We assume without loss of generality that  $t$  follows an HMM  $H$  with a deterministic emission,  $H = (Q_H, \pi_H, \delta_H, \psi_H)$ .

We set  $Q_{H'} = Q_H^{O_\Gamma+1} \times Q$ . Let  $(\pi_{H'}, \delta_{H'})$  be the Markov model on  $Q_{H'}$  such that for all  $d, d' \in Q_H^{O_\Gamma+1}$  and all  $q, q' \in Q$ , we have

$$\pi_{H'}([d, q]) = \begin{cases} p_{(\pi_H, \delta_H)}(d) & \text{if } q = o, \\ 0 & \text{otherwise, and} \end{cases}$$

$$\delta_{H'}([d, q], [d', q']) = \begin{cases} 0 & \text{if } q' \neq \delta(q, \psi_H(d_{\alpha(q)})), \\ 1 & \text{if } q' = \delta(q, \psi_H(d_{\alpha(q)})), d' = d \text{ and } \gamma(q, \psi_H(d_{\alpha(q)})) = 0, \\ p_{(\pi_H, \delta_H)}^*(d, d', \gamma(q, \psi_H(d_{\alpha(q)}))) & \text{if } q' = \delta(q, \psi_H(d_{\alpha(q)})) \text{ and } \gamma(q, \psi_H(d_{\alpha(q)})) > 0, \end{cases}$$

where  $p_{(\pi_H, \delta_H)}^*(d, d', \ell)$  is the probability of observing  $d'$  given that  $d$  occurs  $\ell$  positions before, under the Markov model  $(\pi_H, \delta_H)$ ,

Since the emission function of  $H$  is deterministic, a sequence of hidden states  $z$  of  $Q_H$  determines the emitted text  $t^z = \psi_H(z)$ , which itself determines the sequence  $(\mathbf{q}_\Gamma^{t^z}(i), \mathbf{s}_\Gamma^{t^z}(i))_i$  of pairs state-shift parsed on the input  $(\Gamma, t^z)$ . Let us verify that if  $z$  follows the Markov model  $(\pi_H, \delta_H)$ , then the Markov model  $(\pi_{H'}, \delta_{H'})$  models the sequence  $([\mathbf{d}(i), \mathbf{q}_\Gamma^{t^z}(i)])_i$ , where  $\mathbf{d}(i) = z_{[\mathbf{p}_\Gamma^{t^z}(i), \mathbf{p}_\Gamma^{t^z}(i) + O_\Gamma]}$ .

Under the current assumptions and since the generic algorithm always starts with  $o$ , we have  $\mathbf{q}_\Gamma^{t^z}(0) = o$  and  $P([\mathbf{d}(0), \mathbf{q}_\Gamma^{t^z}(0)]) = p_{(\pi_H, \delta_H)}(z_{[0, O_\Gamma]})$ . The initial state of the sequence  $([\mathbf{d}(i), \mathbf{q}_\Gamma^{t^z}(i)])_i$  does follow the distribution  $\pi_{H'}$ .

Let us assume that, for  $j \geq 0$ , the probability of  $([\mathbf{d}(i), \mathbf{q}_\Gamma^{t^z}(i)])_{[0, j]}$  is

$$p_{(\pi_{H'}, \delta_{H'})} \left( ([\mathbf{d}(i), \mathbf{q}_\Gamma^{t^z}(i)])_{[0, j]} \right).$$

Both the next state  $\mathbf{q}_\Gamma^{t^z}(j+1)$  and the shift  $\mathbf{s}_\Gamma^{t^z}(j)$  only depend on  $\mathbf{q}_\Gamma^{t^z}(j)$  and on the symbol  $x_j = t_{\mathbf{p}_\Gamma^{t^z}(j) + \alpha(\mathbf{q}_\Gamma^{t^z}(j))}^z$ . Both are fully determined by the current state  $[\mathbf{d}(j), \mathbf{q}_\Gamma^{t^z}(j)]$  of  $Q_{H'}$ . In particular, for all  $d \in Q_H^{O_\Gamma+1}$ , if we have  $q' \neq \delta(\mathbf{q}_\Gamma^{t^z}(j), x_j)$  then

$$[\mathbf{d}(j+1), \mathbf{q}_\Gamma^{t^z}(j+1)] \neq [d, q'].$$

If  $\mathbf{s}_\Gamma^{t^z}(j) = 0$  then we have

$$[\mathbf{d}(j+1), \mathbf{q}_\Gamma^{t^z}(j+1)] = [\mathbf{d}(j), \delta(\mathbf{q}_\Gamma^{t^z}(j), x_j)], \quad \text{with probability 1.}$$

Otherwise, by setting  $\mathbf{p}_\Gamma^{t^z}(j+1) = \mathbf{p}_\Gamma^{t^z}(j) + \mathbf{s}_\Gamma^{t^z}(j)$ , we get

$$[\mathbf{d}(j+1), \mathbf{q}_\Gamma^{t^z}(j+1)] = [\mathbf{d}(j+1), \delta(\mathbf{q}_\Gamma^{t^z}(j), x_j)],$$

with probability  $p_{(\pi_H, \delta_H)}^*(\mathbf{d}(j), \mathbf{d}(j+1), \mathbf{s}_\Gamma^{t^z}(j))$ .

Altogether, we get that the probability of  $([\mathbf{d}(i), \mathbf{q}_\Gamma^{t^z}(i)])_{[0, j+1]}$  is equal to

$$p_{(\pi_{H'}, \delta_{H'})} \left( ([\mathbf{d}(i), \mathbf{q}_\Gamma^{t^z}(i)])_{[0, j+1]} \right).$$

The sequence  $([\mathbf{d}(i), \mathbf{q}_\Gamma^{t^z}(i)])_i$  follows the Markov model  $(\pi_{H'}, \delta_{H'})$ . By construction, .  $\square$

Theorem 2 holds for both Markov and iid models and implies that both the sequence of state and the sequence of shifts follow an HMM. If  $t$  follows a Markov model of order  $n$ , one can prove in the same way that the sequence  $(t_{[k_i, k_i + L - 1]}, \mathbf{q}_\Gamma^t(i))_i$  with  $L = \max\{O_\Gamma, n\}$ , follows a Markov model, which may emit the sequence of states and that of shifts. More interestingly, if  $t$  follows an iid model and  $\Gamma$  is non-redundant or standard then the sequence of states parsed on the input  $(\Gamma, t)$  directly follows a Markov model.

**Theorem 3.** Let  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  be a  $w$ -matching machine. If a text  $t$  follows an iid model and  $\Gamma$  is non-redundant (resp. standard) then the sequence of states parsed by the generic algorithm on the input  $(\Gamma, t)$  follows a Markov model  $M = (\pi_M, \delta_M)$ , where for all states  $q$  and  $q'$ ,

- $\pi_M(q) = \begin{cases} 1 & \text{if } q = o, \\ 0 & \text{otherwise;} \end{cases}$
- $\delta_M(q, q') = \sum_{x, \delta(q, x) = q'} \pi(x)$  if  $\Gamma$  is not redundant;
- $\delta_M(q, q') = \frac{\sum_{x, \delta(q, x) = q'} \pi(x)}{\sum_{x, \delta(q, x) \neq \odot} \pi(x)}$  if  $\Gamma$  is standard.

*Proof.* Whatever the text model and the matching machine, the sequence of states always starts with the state  $o$  with probability 1. We have  $\pi_M(o) = 1$  and  $\pi_M(q) = 0$  for all  $q \neq o$ .

If the positions of  $t$  are iid with distribution  $\pi$  and if  $\Gamma$  is non-redundant then the symbols read at each text access are independently drawn from  $\pi$ . It follows that the probability that the state  $q'$  follows the state  $q$  at any iteration is

$$\delta_M(q, q') = \sum_{x, \delta(q, x) = q'} \pi(x),$$

independently of the previous states.

Let us now assume that  $\Gamma$  is standard and that the text  $t$  still follows an iid model  $\pi$ . By construction, the probability  $\delta_M(q, q')$  that the state  $q'$  follows the state  $q$  during the execution of the generic algorithm on the input  $(\Gamma, t)$ , is equal to:

- 1, if there exists a symbol  $x$  such that  $(\alpha(q), x) \in \mathbf{h}_\Gamma(q)$  and  $\delta(q, x) = q'$ ,
- $\sum_{x, \delta(q, x) = q'} \pi(x)$ , otherwise,

independently of the previous states. If there exists a symbol  $x$  such that  $(\alpha(q), x) \in \mathbf{h}_\Gamma(q)$ , then we have  $\delta(q, y) = \odot$  for all symbols  $y \neq x$ . Otherwise, since  $\Gamma$  is valid, there is no symbol  $y$  such that  $\delta(q, y) = \odot$ . In both cases, we have that

$$\delta_M(q, q') = \frac{\sum_{x, \delta(q, x) = q'} \pi(x)}{\sum_{x, \delta(q, x) \neq \odot} \pi(x)}$$

□

## 4.2 Asymptotic speed

In [13, 14], the authors studied the exact distribution of the number of text accesses of some classical algorithms seeking for a pattern in Bernoulli random texts of a given length. We are here rather interested in the asymptotic behavior of algorithms, still in terms of text accesses.

Let  $\mathcal{M}$  be a text model and  $\mathbf{A}$  be an algorithm. The *asymptotic speed* of  $\mathbf{A}$  with respect to  $w$  and under  $\mathcal{M}$  is the limit, when  $n$  goes to infinity, of the expectation of the ratio of  $n$  to the number of text accesses performed by  $\mathbf{A}$  by parsing a text of length  $n$  drawn from  $\mathcal{M}$ . Formally, by putting  $a_{\mathbf{A}}(t)$  for the number of text accesses performed by  $\mathbf{A}$  to parse  $t$ , the asymptotic speed of  $\mathbf{A}$  under  $\mathcal{M}$  is

$$\text{AS}_{\mathcal{M}}(\mathbf{A}) = \lim_{n \rightarrow \infty} \sum_{t \in \mathcal{A}^n} \frac{|t|}{a_{\mathbf{A}}(t)} p_{\mathcal{M}}(t).$$

In order to make the notations less cluttered,  $w$  does not appear neither on  $\text{AS}_{\mathcal{M}}(\mathbf{A})$  nor on  $a_{\mathbf{A}}(t)$ , but these two quantities actually depend on  $w$ . At this point, nothing ensures that the limit above exists.

For all  $w$ -matching machines  $\Gamma$ , we put  $a_{\Gamma}$  for the number of text accesses and  $\text{AS}_{\mathcal{M}}(\Gamma)$  for the asymptotic speed of the generic algorithm with  $\Gamma$  as first input. For a matching machine, the number of text accesses coincides with the number of iterations.

The following remark is a direct consequence of the definition of redundancy and of Remark 4.

**Remark 5.** *If it exists, the asymptotic speed of a non-redundant matching machine is greater than 1.*

In particular, the remark above holds for  $w$ -matching machines which are both standard and compact (Remark 4). It implies that any matching machine can be turned into a matching machine with an asymptotic speed greater than 1 (Proposition 2).

**Lemma 4.** *Let  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  be a  $w$ -matching machine. If  $\Gamma$  is valid then we have for all texts  $t$ ,*

$$\left\lfloor \frac{|t|}{|w|} \right\rfloor \leq a_{\Gamma}(t) \leq (|t| + 1)(|Q| + 1).$$

*Proof.* If there exists a text  $t$  such that  $a_{\Gamma}(t) < \left\lfloor \frac{|t|}{|w|} \right\rfloor$  then there exists  $|w|$  successive positions of  $t$  which are not accessed during the execution of the generic algorithm on the input  $(\Gamma, t)$  [9]. They may contain an occurrence of  $w$  which wouldn't be reported.

If there exists a text  $t$  such that  $a_{\Gamma}(t) > (|t| + 1)(|Q| + 1)$  then there exists an iteration  $i \leq a_{\Gamma}(t) - |Q| - 1$  such that  $s_{\Gamma}^t(j) = 0$  for all  $i \leq j \leq i + |Q|$ . Since there are only  $|Q|$  states, there exist two integers  $k$  and  $\ell$  such that  $i \leq k < \ell \leq i + |Q|$  and  $\mathbf{q}_{\Gamma}^t(k) = \mathbf{q}_{\Gamma}^t(\ell)$ , which contradicts the validity of  $\Gamma$  (Item 2 of Remark 1).  $\square$



We will need the following technical lemma.

**Lemma 5.** *Let  $M = (\pi_M, \delta_M)$  be a Markov model on an alphabet  $\mathcal{Q}_M$  and  $\phi$  be a map from  $\mathcal{Q}_M$  to  $\mathbb{N}$ . Let us assume that we have*

$$\lim_{n \rightarrow \infty} \sum_{i=0}^n \phi(V_i) = \infty \quad \text{with probability 1,}$$

where  $(V_i)_i$  is the Markov chain in which  $V_0$  has the probability distribution  $\pi_M$  and, for all  $i \geq 0$ ,  $P\{V_{i+1} = b \mid V_i = a\} = \delta_M(a, b)$ .

By setting  $S_\kappa(n) = \{v \in \mathcal{Q}_M^* \mid \sum_{i=0}^{|v|-2} \phi(v_i) + \kappa < n \leq \sum_{i=0}^{|v|-1} \phi(v_i) + \kappa\}$  where  $\kappa$  is a non-negative number, the sum

$$\sum_{v \in S_\kappa(n)} \frac{|v|_x}{|v|} p_M(v)$$

converges for all states  $x \in \mathcal{Q}_M$  as  $n$  goes to infinity, to

$$\lim_{k \rightarrow \infty} \sum_{v \in \mathcal{Q}_M^k} \frac{|v|_x}{k} p_M(v).$$

*Proof.* We define the random variable  $F_{x,n}$  as the ratio  $\frac{|V_{[0, \ell_{V,n}-1]}|_x}{\ell_{V,n}}$  where  $\ell_{V,n}$  is the smallest integer such that  $\sum_{i=0}^{\ell_{V,n}-1} \phi(V_i) + \kappa \geq n$ .

Since, under the assumptions of the lemma,  $\lim_{n \rightarrow \infty} \ell_{V,n} = \infty$  with probability 1, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} F_{x,n} &= \lim_{n \rightarrow \infty} \frac{|V_{[0, \ell_{V,n}-1]}|_x}{\ell_{V,n}} \\ &= \lim_{k \rightarrow \infty} \frac{|V_{[0, k-1]}|_x}{k} \quad \text{with probability 1.} \end{aligned} \tag{1}$$

The fact that  $\frac{|V_{[0, k-1]}|_x}{k}$  converges almost surely (a.s.) as  $k$  goes to  $\infty$  is a classical result of Markov chains. In particular, The Ergodic Theorem states that if the chain is irreducible  $\frac{|V_{[0, k-1]}|_x}{k}$  converges a.s. to the probability of the state  $x$  in its stationary distribution [7].

Let us remark that, for all  $v \in \mathcal{Q}_M^*$ , the probability  $P\{\ell_{V,n} = |v| \mid V_{[0, |v|-1]} = v\}$  is 1 if  $v$  verifies  $\sum_{i=0}^{|v|-2} \phi(v_i) + \kappa < n \leq \sum_{i=0}^{|v|-1} \phi(v_i) + \kappa$ , and 0 otherwise. We have, for all  $v \in \mathcal{Q}_M^*$ ,

$$P\{\ell_{V,n} = |v| \text{ and } V_{[0, |v|-1]} = v\} = \begin{cases} p_M(v) & \text{if } v \in S_\kappa(n), \\ 0 & \text{otherwise.} \end{cases}$$

It follows that

$$\begin{aligned} \mathbf{E}(F_{x,n}) &= \sum_{k \geq 0} \left( \sum_{v \in \mathcal{Q}_M^k} \frac{|v|_x}{k} P\{\ell_{V,n} = k \text{ and } V_{[0, k-1]} = v\} \right) \\ &= \sum_{v \in S_\kappa(n)} \frac{|v|_x}{|v|} p_M(v). \end{aligned}$$

Moreover, since  $F_{x,n} \leq 1$ , the bounded convergence theorem gives us that

$$\lim_{n \rightarrow \infty} \mathbf{E}(F_{x,n}) = \mathbf{E}(\lim_{n \rightarrow \infty} F_{x,n}).$$

The sum  $\sum_{v \in S_\kappa(n)} \frac{|v|_x}{|v|} p_M(v)$  does converge as  $n$  goes to  $\infty$ , to

$$\lim_{k \rightarrow \infty} \sum_{v \in \mathcal{Q}_M^k} \frac{|v|_x}{k} p_M(v) \quad (\text{Equation 1}).$$

□

We are now able to prove that the asymptotic speed of a matching machine does exist under an HMM.

**Theorem 4.** *Let  $H$  be a HMM and  $\Gamma$  be a valid matching machine. The sum  $\sum_{t \in \mathcal{A}^n} \frac{|t|}{a_\Gamma(t)} p_H(t)$  converges as  $n$  goes to infinity.*

*Proof.* Let  $H = (Q_H, \pi_H, \delta_H, \phi_H)$  be a HMM and  $t$  be a text. The number of iterations of the generic algorithm on the input  $(\Gamma, t)$  is equal to the number  $a_\Gamma(t)$  of text accesses. From the loop condition of the generic algorithm, we get that

$$\frac{\sum_{i=0}^{a_\Gamma(t)-2} \mathbf{s}_\Gamma^t(i)}{a_\Gamma(t)} + \frac{|w|}{a_\Gamma(t)} < \frac{|t|}{a_\Gamma(t)} \leq \frac{\sum_{i=0}^{a_\Gamma(t)-1} \mathbf{s}_\Gamma^t(i)}{a_\Gamma(t)} + \frac{|w|}{a_\Gamma(t)}. \quad (2)$$

Since the validity of  $\Gamma$  implies that  $\lim_{|t| \rightarrow \infty} a_\Gamma(t) = \infty$  (Lemma 4), Inequality 2 leads to

$$\lim_{n \rightarrow \infty} \sum_{t \in \mathcal{A}^n} \frac{|t|}{a_\Gamma(t)} p_H(t) = \lim_{n \rightarrow \infty} \sum_{t \in \mathcal{A}^n} \frac{\sum_{i=0}^{a_\Gamma(t)-1} \mathbf{s}_\Gamma^t(i)}{a_\Gamma(t)} p_H(t). \quad (3)$$

From Theorem 2, if  $t$  follows the HMM  $H$  then the sequence of shifts  $(\mathbf{s}_\Gamma^t(i))_{0 \leq i < a_\Gamma(t)}$  follows a HMM  $H' = (Q_{H'}, \pi_{H'}, \delta_{H'}, \psi_{H'})$ , which is assumed to have a deterministic emission without loss of generality.  $H' = (Q_{H'}, \pi_{H'}, \delta_{H'}, \phi_{H'})$ . By setting

$$S_\kappa(n) = \left\{ v \in Q_{H'}^* \mid \sum_{i=0}^{|v|-2} \psi_{H'}(v_i) + \kappa < n \leq \sum_{i=0}^{|v|-1} \psi_{H'}(v_i) + \kappa \right\},$$

Equation 3 becomes

$$\begin{aligned} \lim_{n \rightarrow \infty} \sum_{t \in \mathcal{A}^n} \frac{|t|}{a_\Gamma(t)} p_H(t) &= \lim_{n \rightarrow \infty} \sum_{v \in S_{|w|}(n)} \frac{\sum_{i=0}^{|v|-1} \psi_{H'}(v_i)}{|v|} p_{H'}(v) \\ &= \lim_{n \rightarrow \infty} \sum_{v \in S_{|w|}(n)} \sum_{q \in Q_{H'}} \psi_{H'}(q) \frac{|v|_q}{|v|} p_{H'}(v). \end{aligned}$$

Interchanging the order of summation gives us that

$$\sum_{v \in S_{|w|}(n)} \sum_{d \in Q_{H'}} \psi_{H'}(d) \frac{|v|_d}{|v|} p_{H'}(v) = \sum_{d \in Q_{H'}} \psi_{H'}(d) \left( \sum_{v \in S_{|w|}(n)} \frac{|v|_d}{|v|} p_{H'}(v) \right). \quad (4)$$

Since the sequence of shifts follows  $H'$  when the text follows  $H$ , for all  $v \in Q_{H'}^*$ , such that  $p_{(\pi_{H'}, \delta_{H'})}(v) > 0$ , there exists a text  $t$  with  $p_H(t) > 0$  and such that the sequence of shifts parsed on the input  $(\Gamma, t)$  is  $\psi_{H'}(v)$  and  $|v|$  is the number of iterations (or text accesses). Under the assumption that  $\Gamma$  is valid, Lemma 4 implies that

$$\begin{aligned} |v| &\leq (|Q_{H'}| + 1)(|t| + 1) \\ &\leq (|Q_{H'}| + 1) \left( \sum_{i=0}^{|v|} \psi_{H'}(v_i) + 1 \right) \end{aligned}$$

thus

$$\sum_{i=0}^{|v|-1} \psi_{H'}(v_i) \geq \frac{|v|}{|Q_{H'}| + 1} - 1$$

In short, we have  $p_{(\pi_{H'}, \delta_{H'})}(v) > 0 \Rightarrow \sum_{i=0}^{|v|-1} \psi_{H'}(v_i) \geq \beta|v|$  with  $\beta > 0$ , which implies that the Markov model  $(\pi_{H'}, \delta_{H'})$  and the map  $\psi_{H'}$  satisfy the assumptions of Lemma 5. We get that the sum  $\sum_{v \in S_{|w|}(n)} \frac{|v|_q}{|v|} p_{H'}(v)$  converges to a limit frequency  $\alpha_q$  as  $n$  goes to  $\infty$ . From Equation 4, the asymptotic speed  $AS_H(\Gamma)$  does exist and is equal to

$$\sum_{q \in Q_{H'}} \psi_{H'}(q) \alpha_q.$$

□

A more precise result can be stated in the case where the model is Bernoulli and the machine is standard.

**Theorem 5.** *Let  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  be a standard and valid  $w$ -matching machine and  $\pi$  a Bernoulli model. The asymptotic speed of  $\Gamma$  under  $\pi$  is equal to:*

$$AS_\pi(\Gamma) = \sum_{q \in Q} \alpha_q E(q),$$

where  $(\alpha_q)_{q \in Q}$  are the limit frequencies of the states of the Markov model associated to  $\Gamma$  and  $\pi$ , given in Theorem 3 and

$$E(q) = \frac{\sum_{x, \delta(q, x) \neq \odot} \gamma(q, x) \pi_x}{\sum_{x, \delta(q, x) \neq \odot} \pi_x}.$$

*Proof.* Since  $\Gamma$  is standard and valid, Lemma 2 states that any transition from a state  $r$  to a state  $s$  (whatever the symbol read from the text) is associated to a unique shift which will be referred to as  $\phi(r, s)$ .

For all texts  $t$ , we then have

$$\begin{aligned} & \frac{\sum_{i=0}^{a_\Gamma(t)-3} \phi(\mathbf{q}_\Gamma^t(i), \mathbf{q}_\Gamma^t(i+1))}{a_\Gamma(t)} \\ & + \frac{|w|}{a_\Gamma(t)} < \frac{|t|}{a_\Gamma(t)} \leq \frac{\sum_{i=0}^{a_\Gamma(t)-2} \phi(\mathbf{q}_\Gamma^t(i), \mathbf{q}_\Gamma^t(i+1))}{a_\Gamma(t)} + \frac{|w|}{a_\Gamma(t)}. \end{aligned}$$

Theorem 3 tells us that if  $t$  is drawn according  $\pi$  then the sequence  $(\mathbf{q}_\Gamma^t(i))_i$  follows a Markov model  $M = (\pi_M, \delta_M)$ . Let us define

$$S_\kappa(n) = \left\{ v \in Q^* \mid \sum_{i=0}^{|v|-2} \phi(v_i, v_{i+1}) + \kappa < n \leq \sum_{i=0}^{|v|-1} \phi(v_i, v_{i+1}) + \kappa \right\}.$$

From the fact that  $\Gamma$  is valid, we get  $\lim_{|t| \rightarrow \infty} a_\Gamma(t) = \infty$  (Lemma 4) and

$$\lim_{n \rightarrow \infty} \sum_{t \in \mathcal{A}^n} \frac{|t|}{a_\Gamma(t)} \pi(t) = \lim_{n \rightarrow \infty} \sum_{v \in S_\kappa(n)} \frac{\sum_{i=0}^{|v|-2} \phi(v_i, v_{i+1})}{|v|} p_M(v).$$

Basically, we have that

$$\frac{\sum_{i=0}^{|v|-2} \phi(v_i, v_{i+1})}{|v|} = \sum_{d \in Q^2} \phi(d_0, d_1) \frac{|v|_d}{|v|}$$

The sequence  $(v_i v_{i+1})_i$  follows a Markov model with states in  $Q^2$ . The same argument as in the proof of Theorem 4 shows that the assumption of Lemma 5 is granted with  $(v_i v_{i+1})_i$  and  $\phi$ , which gives us that, for all  $d \in Q^2$ ,

$$\begin{aligned} \lim_{n \rightarrow \infty} \sum_{v \in S_\kappa(n)} \frac{|v|_d}{|v|} p_M(v) &= \lim_{k \rightarrow \infty} \sum_{v \in Q^k} \frac{|v|_d}{k} p_M(v) \\ &= \lim_{k \rightarrow \infty} \sum_{v \in Q^k} \frac{|v|_{d_0}}{k} \times \frac{|v|_d}{|v|_{d_0}} p_M(v) \\ &= \alpha_{d_0} \delta_M(d_0, d_1) \end{aligned}$$

Finally we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \sum_{t \in \mathcal{A}^n} \frac{|t|}{a_\Gamma(t)} \pi(t) &= \lim_{n \rightarrow \infty} \sum_{v \in S_\kappa(n)} \frac{\sum_{i=0}^{|v|-2} \phi(v_i, v_{i+1})}{|v|} p_M(v) \\ &= \sum_{d \in Q^2} \phi(d_0, d_1) \lim_{n \rightarrow \infty} \sum_{v \in S_\kappa(n)} \frac{|v|_d}{|v|} p_M(v) \\ &= \sum_{d \in Q^2} \phi(d_0, d_1) \alpha_{d_0} \delta_M(d_0, d_1) \end{aligned}$$

$$= \sum_{d_0 \in Q} \alpha_{d_0} \sum_{d_1 \in Q} \phi(d_0, d_1) \delta_M(d_0, d_1)$$

With Theorem 3, we have that

$$\begin{aligned} \sum_{d_1 \in Q} \phi(d_0, d_1) \delta_M(d_0, d_1) &= \sum_{d_1 \in Q} \phi(d_0, d_1) \frac{\sum_{x, \delta(d_0, x) = d_1} \pi(x)}{\sum_{x, \delta(d_0, x) \neq \odot} \pi(x)} \\ &= E(d_0) \end{aligned}$$

□

## 5 Withdrawing inefficient states

We shall see that some states of a matching machine may be removed without decreasing its asymptotic speed under a given iid model.

### 5.1 Redirecting transitions

**Theorem 6.** *Let  $\pi$  be an iid model,  $t$  a text drawn from  $\pi$ ,  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  be a  $w$ -matching machine and  $\dot{q}$  and  $\ddot{q}$  be two states which are such that, under the notations of Section 3.2,*

- *for all states  $r$  and all symbols  $x$  and  $y$ ,  $\delta(r, x) = \delta(r, y) \Rightarrow \gamma(r, x) = \gamma(r, y)$ ;*
- *the sequence of states of the generic algorithm on the input  $(\Gamma, t)$  follows a Markov model  $M = (\pi_M, \delta_M)$ ;*
- *the sequence of states of the generic algorithm on the input  $(\Gamma_{\dot{q} \triangleright \ddot{q}}, t)$  follows a Markov model  $\dot{M} = (\pi_{\dot{M}}, \delta_{\dot{M}})$  which is such that*
  - *if  $\ddot{q} \neq o$  then  $\pi_{\dot{M}} = \pi_M$ , otherwise  $\pi_{\dot{M}}(\dot{q}) = 1$  and  $\pi_{\dot{M}}(s) = 0$  for all states  $s \neq \dot{q}$ ,*
  - *$\delta_{\dot{M}}(r, s) = \delta_M(r, s)$  for all states  $s \neq \dot{q}$ ,*
  - *$\delta_{\dot{M}}(r, \dot{q}) = \delta_M(r, \dot{q}) + \delta_M(r, \ddot{q})$ ;*
- *the sequence of states of the generic algorithm on the input  $(\Gamma_{\dot{q} \triangleright \ddot{q}}, t)$  follows a Markov model  $\ddot{M} = (\pi_{\ddot{M}}, \delta_{\ddot{M}})$  which is such that*
  - *if  $\dot{q} \neq o$  then  $\pi_{\ddot{M}} = \pi_M$ , otherwise  $\pi_{\ddot{M}}(\ddot{q}) = 1$  and  $\pi_{\ddot{M}}(s) = 0$  for all states  $s \neq \ddot{q}$ ,*
  - *$\delta_{\ddot{M}}(r, s) = \delta_M(r, s)$  for all states  $s \neq \ddot{q}$ ,*
  - *$\delta_{\ddot{M}}(r, \ddot{q}) = \delta_M(r, \dot{q}) + \delta_M(r, \ddot{q})$ .*

We have

$$\text{AS}_\pi(\Gamma) \leq \max\{\text{AS}_\pi(\Gamma_{\dot{q} \triangleright \dot{q}}), \text{AS}_\pi(\Gamma_{\dot{q} \triangleright \dot{q}})\}.$$

*Proof.* Under the assumptions of the theorem, the sequence  $(\mathbf{q}_\Gamma^t(i))_i$  follows a Markov model  $M = (\pi_M, \delta_M)$  and any transition from a state  $r$  to a state  $s$  (whatever the symbol read from the text) is associated to a unique shift which will be referred to as  $\phi(r, s)$ .

By defining the set  $S_\kappa(n)$  as

$$S_\kappa(n) = \{v \in Q^* \mid \sum_{i=0}^{|v|-3} \phi(v_i, v_{i+1}) + \kappa < n \leq \sum_{i=0}^{|v|-2} \phi(v_i, v_{i+1}) + \kappa\},$$

we have

$$\begin{aligned} \text{AS}_\pi(\Gamma) &= \lim_{n \rightarrow \infty} \sum_{t \in \mathcal{A}^n} \frac{|t|}{a_\Gamma(t)} p_\pi(t) \\ &= \lim_{n \rightarrow \infty} \sum_{v \in S_{|w|}(n)} \frac{\sum_{i=0}^{|v|-2} \phi(v_i, v_{i+1})}{|v|} p_M(v). \end{aligned}$$

A similar argument as that of the proof of Lemma 5 shows that

$$\text{AS}_\pi(\Gamma) = \lim_{k \rightarrow \infty} \sum_{v \in Q^k} \frac{\sum_{i=0}^{|v|-2} \phi(v_i, v_{i+1})}{|v|} p_M(v). \quad (5)$$

Let now consider the Markov chain  $V = (V_i)_i$  where  $V_0 = o$  and, for all  $i \geq 0$ ,  $P\{V_{i+1} = s \mid V_i = r\} = \delta_M(r, s)$ . The chain  $V$  models the execution process of the generic algorithm on the input  $(\Gamma, t)$  with  $t$  iid. Let us rewrite Equation 5 as

$$\text{AS}_\pi(\Gamma) = \lim_{k \rightarrow \infty} \mathbf{E} \left( \frac{\sum_{i=0}^{k-1} \phi(V_i, V_{i+1})}{k} \right).$$

The set of states of  $V$  (or of any Markov chain) may be partitioned in a unique way into the class  $\mathcal{T}$  of its transient states, which may be empty, and a positive number  $c$  of non-empty recurrent classes (i.e. closed communicating classes)  $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_c$ .

For all  $1 \leq m \leq c$ , we define the Markov chains  $V^{(m)} = (V_i^{(m)})_i$  where  $V_0^{(m)} \in \mathcal{C}_m$  and, for all  $i \geq 0$ ,  $P\{V_{i+1} = s \mid V_i = r\} = \delta_M(r, s)$ . All the chains  $V^{(m)}$  are irreducible. The asymptotic speed of the class  $\mathcal{C}_m$  is noted  $\text{AS}_\pi^{(m)}(\Gamma)$  and defined as

$$\text{AS}_\pi^{(m)}(\Gamma) = \lim_{k \rightarrow \infty} \mathbf{E} \left( \frac{\sum_{i=0}^{k-1} \phi(V_i^{(m)}, V_{i+1}^{(m)})}{k} \right).$$

For two subsets  $\mathcal{E}$  and  $\mathcal{F}$  of  $Q$  and  $r$  a state of  $Q \setminus \mathcal{E}$ , let  $f_\mathcal{E}^{(n)}(r, \mathcal{F})$  be the probability for the random variable  $V_n$  to be in  $\mathcal{F}$  without visiting any state of  $\mathcal{E} \cup \mathcal{F}$  from 1 to  $n-1$ , being given that  $V_0 = r$ , namely

$$f_\mathcal{E}^{(n)}(r, \mathcal{F}) = P\{V_n \in \mathcal{F} \text{ and } V_k \notin \mathcal{E} \cup \mathcal{F} \text{ for all } 0 < k < n \mid V_0 = r\}.$$

We also define

$$f_{\mathcal{E}}(r, \mathcal{F}) = \sum_{n=1}^{\infty} f_{\mathcal{E}}^{(n)}(r, \mathcal{F}).$$

Starting from  $o$  (or any state), the chain  $V$  may visit some transient states but goes to one or another recurrent class in a time which is a.s. finite. Then, it stays in this recurrent class indefinitely. By writing  $f(r, \mathcal{F})$  for  $f_{\emptyset}(r, \mathcal{F})$ , we have

$$\text{AS}_{\pi}(\Gamma) = \sum_{m=1}^c f(o, \mathcal{C}_m) \text{AS}_{\pi}^{(m)}(\Gamma). \quad (6)$$

Since the chain  $V$  ends up in a recurrent class with probability 1, the law of total probability gives us that

$$\sum_{m=1}^c f(o, \mathcal{C}_m) = 1.$$

In order to prove the inequality of the theorem, we have to distinguish different cases according to which classes the states  $\dot{q}$  and  $\ddot{q}$  belong.

#### Case 1 – $\dot{q}$ and $\ddot{q}$ are both transient

Since  $\dot{q}$  and  $\ddot{q}$  are reachable (from our implicit assumption), the state  $o$ , which leads to  $\dot{q}$  and  $\ddot{q}$ , is transient as well.

For all subsets  $\mathcal{S} \subset Q$  and all states  $r \in \mathcal{S}$ , let  $g_{\mathcal{S}}^{(n)}(r)$  denote the probability that  $V_n = r$  is the last state of  $\mathcal{S}$  occurring in  $V$ , conditioned on  $V_0 \in \mathcal{S}$ , namely

$$g_{\mathcal{S}}^{(n)}(r) = \text{P}\{V_n = r \text{ and } V_k \notin \mathcal{S} \text{ for all } k > n \mid V_0 \in \mathcal{S}\}.$$

We also define

$$g_{\mathcal{S}}(r) = \sum_{n=1}^{\infty} g_{\mathcal{S}}^{(n)}(r).$$

Let us remark that if  $\mathcal{S}$  contains any recurrent state reachable from  $r$  then both  $g_{\mathcal{S}}^{(n)}(r)$  and  $g_{\mathcal{S}}(r)$  are zero.

We have

$$\begin{aligned} f(o, \mathcal{C}_m) &= f_{\{\dot{q}, \ddot{q}\}}(o, \mathcal{C}_m) \\ &\quad + f(o, \{\dot{q}, \ddot{q}\}) [g_{\{\dot{q}, \ddot{q}\}}(\dot{q}) f_{\{\dot{q}, \ddot{q}\}}(\dot{q}, \mathcal{C}_m) + g_{\{\dot{q}, \ddot{q}\}}(\ddot{q}) f_{\{\dot{q}, \ddot{q}\}}(\ddot{q}, \mathcal{C}_m)]. \end{aligned} \quad (7)$$

Substituting the coefficients of Equation 7 in Equation 6, gives us

$$\begin{aligned}
\text{AS}_\pi(\Gamma) &= \sum_{m=1}^c (f_{\{\dot{q}, \ddot{q}\}}(o, \mathcal{C}_m) + f(o, \{\dot{q}, \ddot{q}\}) [g_{\{\dot{q}, \ddot{q}\}}(\dot{q}) f_{\{\dot{q}, \ddot{q}\}}(\dot{q}, \mathcal{C}_m) \\
&\quad + g_{\{\dot{q}, \ddot{q}\}}(\ddot{q}) f_{\{\dot{q}, \ddot{q}\}}(\ddot{q}, \mathcal{C}_m)]) \text{AS}_\pi^{(m)}(\Gamma) \\
&= \sum_{m=1}^c f_{\{\dot{q}, \ddot{q}\}}(o, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) \\
&\quad + f(o, \{\dot{q}, \ddot{q}\}) \left[ g_{\{\dot{q}, \ddot{q}\}}(\dot{q}) \sum_{m=1}^c f_{\{\dot{q}, \ddot{q}\}}(\dot{q}, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) \right. \\
&\quad \left. + g_{\{\dot{q}, \ddot{q}\}}(\ddot{q}) \sum_{m=1}^c f_{\{\dot{q}, \ddot{q}\}}(\ddot{q}, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) \right].
\end{aligned}$$

Since both  $\dot{q}$  and  $\ddot{q}$  are transient, there exists almost surely an integer  $n$  such that neither  $\dot{q}$  nor  $\ddot{q}$  occurs after  $n$ . Altogether with the law of total probability, this implies that

$$g_{\{\dot{q}, \ddot{q}\}}(\dot{q}) + g_{\{\dot{q}, \ddot{q}\}}(\ddot{q}) = 1. \quad (8)$$

Let now consider the matching machines  $\Gamma_{\ddot{q} \triangleright \dot{q}}$  and  $\Gamma_{\dot{q} \triangleright \ddot{q}}$ . Under the assumptions of the theorem, the states of the generic algorithm follows the Markov models  $\dot{M}$  and  $\ddot{M}$  on the inputs  $(\Gamma_{\ddot{q} \triangleright \dot{q}}, t)$  and  $(\Gamma_{\dot{q} \triangleright \ddot{q}}, t)$ . Still from these assumptions, the models  $\dot{M}$  and  $\ddot{M}$  differ with  $M$  only in the probability transitions ending on  $\dot{q}$  and on  $\ddot{q}$  respectively. By putting  $\dot{f}$  and  $\ddot{f}$  for the analogs of  $f$  with  $\dot{M}$  and  $\ddot{M}$  respectively, we have, for all  $1 \leq m \leq c$ ,

- $\text{AS}_\pi^{(m)}(\Gamma_{\ddot{q} \triangleright \dot{q}}) = \text{AS}_\pi^{(m)}(\Gamma_{\dot{q} \triangleright \ddot{q}}) = \text{AS}_\pi^{(m)}(\Gamma)$ ,
- $f_{\{\dot{q}, \ddot{q}\}}(o, \mathcal{C}_m) = \dot{f}_{\dot{q}}(o, \mathcal{C}_m) = \ddot{f}_{\ddot{q}}(o, \mathcal{C}_m)$ ,
- $f(o, \{\dot{q}, \ddot{q}\}) = \dot{f}(o, \dot{q}) = \ddot{f}(o, \ddot{q})$ ,
- $f_{\{\dot{q}, \ddot{q}\}}(\dot{q}, \mathcal{C}_m) = \dot{f}_{\dot{q}}(\dot{q}, \mathcal{C}_m)$ ,
- $f_{\{\dot{q}, \ddot{q}\}}(\ddot{q}, \mathcal{C}_m) = \ddot{f}_{\ddot{q}}(\ddot{q}, \mathcal{C}_m)$ .

It follows that

$$\begin{aligned}
\dot{f}(o, \mathcal{C}_m) &= \dot{f}_{\dot{q}}(o, \mathcal{C}_m) + \dot{f}(o, \dot{q}) \dot{f}_{\dot{q}}(\dot{q}, \mathcal{C}_m) \\
&= f_{\{\dot{q}, \ddot{q}\}}(o, \mathcal{C}_m) + f(o, \{\dot{q}, \ddot{q}\}) f_{\{\dot{q}, \ddot{q}\}}(\dot{q}, \mathcal{C}_m)
\end{aligned}$$

and

$$\begin{aligned}
\ddot{f}(o, \mathcal{C}_m) &= \ddot{f}_{\ddot{q}}(o, \mathcal{C}_m) + \ddot{f}(o, \ddot{q}) \ddot{f}_{\ddot{q}}(\ddot{q}, \mathcal{C}_m) \\
&= f_{\{\dot{q}, \ddot{q}\}}(o, \mathcal{C}_m) + f(o, \{\dot{q}, \ddot{q}\}) f_{\{\dot{q}, \ddot{q}\}}(\ddot{q}, \mathcal{C}_m).
\end{aligned}$$



Considering Equation 6 for the asymptotic speeds of  $\Gamma_{\dot{q} \triangleright \dot{q}}$  and  $\Gamma_{\dot{q} \triangleright \ddot{q}}$  and the relations just above, leads to

$$\begin{aligned} \text{AS}_\pi(\Gamma_{\dot{q} \triangleright \dot{q}}) &= \sum_{m=1}^c \dot{f}(o, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma_{\dot{q} \triangleright \dot{q}}) \\ &= \sum_{m=1}^c f_{\{\dot{q}, \dot{q}\}}(o, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) + f(o, \{\dot{q}, \ddot{q}\}) \sum_{m=1}^c f_{\{\dot{q}, \ddot{q}\}}(\dot{q}, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) \end{aligned}$$

and

$$\begin{aligned} \text{AS}_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}}) &= \sum_{m=1}^c \ddot{f}(o, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma_{\dot{q} \triangleright \ddot{q}}) \\ &= \sum_{m=1}^c f_{\{\dot{q}, \ddot{q}\}}(o, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) + f(o, \{\dot{q}, \ddot{q}\}) \sum_{m=1}^c f_{\{\dot{q}, \ddot{q}\}}(\ddot{q}, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma). \end{aligned}$$

Since a convex combination is smaller than the greatest of its elements, Equation 8 gives us

$$\begin{aligned} &g_{\{\dot{q}, \ddot{q}\}}(\dot{q}) \sum_{m=1}^c f_{\{\dot{q}, \ddot{q}\}}(\dot{q}, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) + g_{\{\dot{q}, \ddot{q}\}}(\ddot{q}) \sum_{m=1}^c f_{\{\dot{q}, \ddot{q}\}}(\ddot{q}, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) \\ &\leq \max \left\{ \sum_{m=1}^c f_{\{\dot{q}, \ddot{q}\}}(\dot{q}, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma), \sum_{m=1}^c f_{\{\dot{q}, \ddot{q}\}}(\ddot{q}, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) \right\} \end{aligned}$$

and, finally,

$$\text{AS}_\pi(\Gamma) \leq \max \{ \text{AS}_\pi(\Gamma_{\dot{q} \triangleright \dot{q}}), \text{AS}_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}}) \}.$$

### Case 2 – $\dot{q}$ is transient and $\ddot{q}$ is recurrent

Let  $\mathcal{C}_k$  be the recurrent class to which  $\ddot{q}$  belongs. We distinguish between two sub-cases according to whether or not  $\mathcal{C}_k$  is the only recurrent class reachable from  $\ddot{q}$ .

**Case 2a** –  $f(\dot{q}, \mathcal{C}_k) < 1$  It implies that  $\dot{q}$  leads to a recurrent class  $\mathcal{C}_\ell$  with  $\ell \neq k$ . Let us consider the Markov chain  $\dot{V} = (\dot{V}_i)_i$  where  $\dot{V}_0 = o$  and, for all  $i \geq 0$ ,  $P\{\dot{V}_{i+1} = s \mid \dot{V}_i = r\} = \delta_{\dot{M}}(r, s)$ . The set of states of  $\dot{V}$  is  $Q \setminus \{\ddot{q}\}$ . Redirecting all the transitions that end with  $\ddot{q}$ , to  $\dot{q}$  makes all the states of  $\mathcal{C}_k \setminus \{\ddot{q}\}$  transient in  $\dot{V}$ . In particular, the part of the asymptotic speed which comes from  $\mathcal{C}_k$  in  $\text{AS}_\pi(\Gamma)$  just vanishes in  $\text{AS}_\pi(\Gamma_{\dot{q} \triangleright \dot{q}})$ . For all  $m \neq k$ , the different ways of reaching  $\mathcal{C}_m$  from  $\dot{q}$  in the chain  $\dot{V}$  may be split into the ways which follows a redirected transition and the ways which don't. Under the theorem's assumptions, any path from  $\dot{q}$  to  $\mathcal{C}_m$  in  $\dot{V}$  which contains no redirected transition has the same probability as in the chain  $V$ . Reciprocally, a path from  $\dot{q}$  to  $\mathcal{C}_m$  in  $V$  contains no state of  $\mathcal{C}_k$ , thus no transition which is redirected in  $\dot{V}$ . In other words, the probability of reaching  $\mathcal{C}_m$  in  $\dot{V}$  without following a redirected transition being given that we start at  $\dot{q}$ , is exactly  $f(\dot{q}, \mathcal{C}_m)$ .

On the other hand, since  $\ddot{q} \in \mathcal{C}_k$  and  $\mathcal{C}_k$  is a recurrent class of  $V$ , the probability of following at least a redirected transition in a path being given that the path starts from  $\dot{q}$ , is  $f(\dot{q}, \mathcal{C}_k)$ . Moreover, since all the redirected transitions end at  $\dot{q}$ , we have that

$$\dot{f}(\dot{q}, \mathcal{C}_m) = f(\dot{q}, \mathcal{C}_m) + f(\dot{q}, \mathcal{C}_k)\dot{f}(\dot{q}, \mathcal{C}_m), \quad \text{thus}$$

$$\dot{f}(\dot{q}, \mathcal{C}_m) = \frac{f(\dot{q}, \mathcal{C}_m)}{1 - f(\dot{q}, \mathcal{C}_k)} = \frac{f(\dot{q}, \mathcal{C}_m)}{\sum_{\substack{\ell=1 \\ \ell \neq k}}^c f(\dot{q}, \mathcal{C}_\ell)}, \quad \text{for all } m \neq k.$$

For all  $1 \leq m \leq c$  with  $m \neq k$  and since the paths of  $\mathcal{C}_m$  and those from  $o$  to  $\dot{q}$  or to a state of  $\mathcal{C}_m$ , never visit  $\ddot{q}$ , we have

- $\text{AS}_\pi^{(m)}(\Gamma_{\ddot{q} \triangleright \dot{q}}) = \text{AS}_\pi^{(m)}(\Gamma)$ ,
- $\dot{f}_{\dot{q}}(o, \mathcal{C}_m) = f_{\dot{q}}(o, \mathcal{C}_m)$ ,
- $\dot{f}(o, \dot{q}) = f(o, \dot{q})$ .

Conversely, redirecting all the transitions that end with  $\dot{q}$ , to  $\ddot{q}$  increases the part of the asymptotic speed which comes from  $\mathcal{C}_k$ . We have, for all  $1 \leq m \leq c$ ,

- $\text{AS}_\pi^{(m)}(\Gamma_{\dot{q} \triangleright \ddot{q}}) = \text{AS}_\pi^{(m)}(\Gamma)$ ,
- $\ddot{f}(o, \mathcal{C}_m) = f(o, \mathcal{C}_m) - f_{\dot{q}}(o, \mathcal{C}_m)$  if  $m \neq k$ ,
- $\ddot{f}(o, \mathcal{C}_k) = f(o, \mathcal{C}_k) + \sum_{\substack{\ell=1 \\ \ell \neq k}}^c f(\dot{q}, \mathcal{C}_\ell)$ .

For all  $1 \leq m \leq c$ , we have

$$f(o, \mathcal{C}_m) = f_{\dot{q}}(o, \mathcal{C}_m) + f(o, \dot{q})f(\dot{q}, \mathcal{C}_m)$$

and, for  $1 \leq m \leq c$  with  $m \neq k$ ,

$$\dot{f}(o, \mathcal{C}_m) = \dot{f}_{\dot{q}}(o, \mathcal{C}_m) + \dot{f}(o, \dot{q})\dot{f}(\dot{q}, \mathcal{C}_m).$$

With Equation 6, it implies that

$$\text{AS}_\pi(\Gamma) = \sum_{m=1}^c f_{\dot{q}}(o, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) + f(o, \dot{q}) \sum_{m=1}^c f(\dot{q}, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma), \quad (9)$$

$$\text{AS}_\pi(\Gamma_{\ddot{q} \triangleright \dot{q}}) = \sum_{m=1}^c f_{\dot{q}}(o, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) + f(o, \dot{q}) \frac{\sum_{\substack{\ell=1 \\ \ell \neq k}}^c f(\dot{q}, \mathcal{C}_\ell) \text{AS}_\pi^{(m)}(\Gamma)}{\sum_{\substack{\ell=1 \\ \ell \neq k}}^c f(\dot{q}, \mathcal{C}_\ell)}, \quad (10)$$

$$\text{AS}_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}}) = \sum_{m=1}^c f_{\dot{q}}(o, \mathcal{C}_m) \text{AS}_\pi^{(m)}(\Gamma) + f(o, \dot{q}) \text{AS}_\pi^{(k)}(\Gamma). \quad (11)$$

We recall that  $\sum_{m=1}^c f(\dot{q}, \mathcal{C}_m) = 1$ . From Equations 9, 10 and 11, we get that

- $AS_\pi(\Gamma) \geq AS_\pi(\Gamma_{\ddot{q} \triangleright \dot{q}})$  if and only if  $AS_\pi^{(k)}(\Gamma) \leq \frac{\sum_{\substack{\ell=1 \\ \ell \neq k}}^c f(\dot{q}, \mathcal{C}_m) AS_\pi^{(m)}(\Gamma)}{\sum_{\substack{\ell=1 \\ \ell \neq k}}^c f(\dot{q}, \mathcal{C}_m)}$ ,
- $AS_\pi(\Gamma) \leq AS_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}})$  if and only if  $AS_\pi^{(k)}(\Gamma) \geq \frac{\sum_{\substack{\ell=1 \\ \ell \neq k}}^c f(\dot{q}, \mathcal{C}_m) AS_\pi^{(m)}(\Gamma)}{\sum_{\substack{\ell=1 \\ \ell \neq k}}^c f(\dot{q}, \mathcal{C}_m)}$ .

In all cases, we have

$$AS_\pi(\Gamma) \leq \max\{AS_\pi(\Gamma_{\ddot{q} \triangleright \dot{q}}), AS_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}})\}.$$

**Case 2b** –  $f(\dot{q}, \mathcal{C}_k) = 1$  It means that  $\dot{q}$  leads only to the recurrent class  $\mathcal{C}_k$ . Redirecting all the transitions that end with  $\ddot{q}$ , to  $\dot{q}$  just replaces the recurrent class  $\mathcal{C}_k$  of  $M$ , by the recurrent class  $\dot{\mathcal{C}}_k$  of  $\dot{M}$ , in which  $\ddot{q}$  plays the role of  $\dot{q}$ . Let  $AS_\pi^{(k)}(\Gamma_{\ddot{q} \triangleright \dot{q}})$  be the asymptotic speed of  $\dot{\mathcal{C}}_k$ . We remark, first, that  $\dot{f}(o, \dot{\mathcal{C}}_k) = f(o, \mathcal{C}_k)$  and, second, that, for all  $m \neq k$ ,  $AS_\pi^{(m)}(\Gamma_{\ddot{q} \triangleright \dot{q}}) = AS_\pi^{(m)}(\Gamma)$ .

Conversely, redirecting all the transitions that end with  $\dot{q}$ , to  $\ddot{q}$  does not change any recurrent class between  $M$  and  $\ddot{M}$ . Moreover we have  $\dot{f}(o, \mathcal{C}_k) = f(o, \mathcal{C}_k)$  and, more generally,  $\dot{f}(o, \mathcal{C}_m) = f(o, \mathcal{C}_m)$  for all  $1 \leq m \leq c$ . With Equation 6, we get that  $AS_\pi(\Gamma) = AS_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}})$ .

In short, we have  $AS_\pi(\Gamma_{\ddot{q} \triangleright \dot{q}}) \geq AS_\pi(\Gamma) = AS_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}})$  if and only if  $AS_\pi^{(k)}(\Gamma_{\ddot{q} \triangleright \dot{q}}) \geq AS_\pi^{(k)}(\Gamma)$ , which leads to

$$AS_\pi(\Gamma) \leq \max\{AS_\pi(\Gamma_{\ddot{q} \triangleright \dot{q}}), AS_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}})\}.$$

### Case 3 – $\dot{q}$ is recurrent and $\ddot{q}$ is transient

This case is perfectly symmetrical with Case 2.

### Case 4 – $\dot{q}$ and $\ddot{q}$ are both recurrent

We have to distinguish between two sub-cases according to whether  $\dot{q}$  and  $\ddot{q}$  are in the same recurrent class.

**Case 4a** –  $\dot{q}$  and  $\ddot{q}$  are in two different recurrent classes Let  $\mathcal{C}_k$  be the class of  $\dot{q}$  and  $\mathcal{C}_\ell$  be the class of  $\ddot{q}$ . Redirecting all the transitions that end with  $\ddot{q}$ , to  $\dot{q}$  makes all the states of  $\mathcal{C}_\ell$  transient (i.e.  $\mathcal{C}_\ell$  is not a recurrent class of  $\dot{M}$ ). Moreover, since all the states of  $\mathcal{C}_\ell$  lead to  $\dot{q} \in \mathcal{C}_k$  in  $\dot{M}$ , we have  $\dot{f}(o, \mathcal{C}_k) = f(o, \mathcal{C}_k) + f(o, \mathcal{C}_\ell)$  and  $\dot{f}(o, \mathcal{C}_m) = f(o, \mathcal{C}_m)$  for all  $m$  different from both  $k$  and  $\ell$ . Redirecting all the transitions that end with  $\dot{q}$ , to  $\ddot{q}$  leads to symmetrical considerations. It follows that we have

- $AS_\pi(\Gamma_{\ddot{q} \triangleright \dot{q}}) \geq AS_\pi(\Gamma)$  if and only if  $AS_\pi^{(k)}(\Gamma) \geq AS_\pi^{(\ell)}(\Gamma)$ ,
- $AS_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}}) \geq AS_\pi(\Gamma)$  if and only if  $AS_\pi^{(k)}(\Gamma) \leq AS_\pi^{(\ell)}(\Gamma)$ .

We get again

$$AS_\pi(\Gamma) \leq \max\{AS_\pi(\Gamma_{\ddot{q} \triangleright \dot{q}}), AS_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}})\}.$$

**Case 4b –  $\dot{q}$  and  $\ddot{q}$  belong to a same recurrent class  $\mathcal{C}_k$**  Redirecting transitions toward  $\dot{q}$  or  $\ddot{q}$  does not change neither the asymptotic speeds of the recurrent classes  $(\mathcal{C}_m)_{m \neq k}$ , nor the probabilities to end up in one of these classes from  $o$ . We start by focusing on the class  $\mathcal{C}_k$ .

Since  $V^{(k)}$  is irreducible, assuming that  $V_0^{(k)} = \dot{q}$  is convenient and does not influence  $\text{AS}_\pi^{(k)}(\Gamma)$ .

Let us define  $A_n$  as the position of the  $n^{\text{th}}$  occurrence of  $\dot{q}$  or  $\ddot{q}$  in  $V^{(k)}$ . Namely  $(A_n)_n$  is such that  $A_0 = 0$  (since we assume  $V_0^{(k)} = \dot{q}$ ) and for all  $n \geq 0$ ,

- $V_{A_n}^{(k)} \in \{\dot{q}, \ddot{q}\}$ ,
- for all  $A_n < i < A_{n+1}$ ,  $V_i^{(k)} \notin \{\dot{q}, \ddot{q}\}$ .

Let  $I_i^{\dot{q}}$  (resp.  $I_i^{\ddot{q}}$ ) be such that  $A_{I_i^{\dot{q}}}$  (resp.  $A_{I_i^{\ddot{q}}}$ ) is the position of the  $i^{\text{th}}$  occurrence of  $\dot{q}$  (resp. of  $\ddot{q}$ ) in  $V^{(k)}$ . For all positions  $i$ , we put  $N_i^{\dot{q}}$  (resp.  $N_i^{\ddot{q}}$ ) for the number of occurrences of  $\dot{q}$  (resp. of  $\ddot{q}$ ) in  $V_0^{(k)}, \dots, V_i^{(k)}$ . We set  $N_i = N_i^{\dot{q}} + N_i^{\ddot{q}}$  and we define the binary random variables  $F_i^{\dot{q}}$  and  $F_i^{\ddot{q}}$  as

$$F_i^{\dot{q}} = \begin{cases} 1 & \text{if } I_{N_i^{\dot{q}}}^{\dot{q}} > I_{N_i^{\ddot{q}}}^{\ddot{q}}, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and } F_i^{\ddot{q}} = \begin{cases} 1 & \text{if } I_{N_i^{\dot{q}}}^{\dot{q}} < I_{N_i^{\ddot{q}}}^{\ddot{q}}, \\ 0 & \text{otherwise.} \end{cases}$$

By setting  $D_i = \sum_{j=A_i}^{A_{i+1}-1} \phi(V_j^{(k)}, V_{j+1}^{(k)})$  we get, for all integers  $n$ ,

$$\frac{\sum_{i=0}^{N_n-1} D_i}{n} \leq \frac{\sum_{j=0}^n \phi(V_j^{(k)}, V_{j+1}^{(k)})}{n} \leq \frac{\sum_{i=0}^{N_n} D_i}{n}.$$

Decomposing the sums above leads to

$$\frac{\sum_{i=0}^{N_n^{\dot{q}}-F_n^{\dot{q}}} D_{I_i^{\dot{q}}}}{n} + \frac{\sum_{i=0}^{N_n^{\ddot{q}}-F_n^{\ddot{q}}} D_{I_i^{\ddot{q}}}}{n} \leq \frac{\sum_{j=0}^n \phi(V_j^{(k)}, V_{j+1}^{(k)})}{n} \leq \frac{\sum_{i=0}^{N_n^{\dot{q}}} D_{I_i^{\dot{q}}}}{n} + \frac{\sum_{i=0}^{N_n^{\ddot{q}}} D_{I_i^{\ddot{q}}}}{n}.$$

Let now consider  $\Gamma_{\dot{q}, \ddot{q}}$  and the corresponding Markov model  $\dot{M}$ . We define the Markov chain  $\dot{V} = (\dot{V}_i)_i$  with  $\dot{V}_0 = o$  and, for all  $i \geq 0$ ,  $\text{P}\{\dot{V}_{i+1} = s \mid \dot{V}_i = r\} = \delta_{\dot{M}}(r, s)$ . By construction, the chain  $\dot{V}$  contains all the recurrent classes  $(\mathcal{C}_m)_{m \neq k}$ . Moreover, since  $\dot{q}$  and  $\ddot{q}$  are both in a recurrent class of  $V$ , all the states which were transient with  $V$  are still transient in  $\dot{V}$ . In particular, for all  $m \neq k$ , no path from  $o$  to a recurrent class  $\mathcal{C}_m$  contains a redirected transition. We have

$$\begin{aligned} f(o, \mathcal{C}_m) &= f_{\{\dot{q}, \ddot{q}\}}(o, \mathcal{C}_m) \\ &= \dot{f}_{\{\dot{q}, \ddot{q}\}}(o, \mathcal{C}_m) \\ &= \dot{f}(o, \mathcal{C}_m). \end{aligned}$$

Since all the states that lead to  $\dot{q}$  in the chain  $V$ , still lead to  $\dot{q}$  in the chain  $\dot{V}$ ,  $\dot{V}$  contains a recurrent class  $\dot{\mathcal{C}}_k$  to which  $\dot{q}$  belongs. Let  $q \neq \dot{q}$  be a state of  $\mathcal{C}_k$ . Several possibilities arise:

- if  $q$  is reachable from  $\dot{q}$  in  $\dot{V}$  then  $q \in \dot{\mathcal{C}}_k$ ;
- if  $q$  is reachable from  $o$  but not from  $\dot{q}$  then  $q$  is transient in  $\dot{V}$ ;
- if  $q$  is not reachable from  $o$  then it is not a state of  $\dot{V}$ .

In short, the chain  $\dot{V}$  contains all the recurrent classes  $(\mathcal{C}_m)_{m \neq k}$ , a non-empty recurrent class  $\dot{\mathcal{C}}_k \subset \mathcal{C}_k$ , a set of transient states which contains that of  $V$ . We have

$$\dot{f}(o, \dot{\mathcal{C}}_k) = f(o, \mathcal{C}_k).$$

By defining, for all  $i \geq 0$ ,

- $\dot{A}_i$  as the position of the  $i^{\text{th}}$  occurrence of  $\dot{q}$  in  $\dot{V}^{(k)}$ ,
- $\dot{N}_i^{\dot{q}}$  as the number of occurrences of  $\dot{q}$  in  $\dot{V}_0^{(k)}, \dots, \dot{V}_i^{(k)}$ ,
- $\dot{D}_i$  as  $\dot{D}_i = \sum_{j=\dot{A}_i}^{\dot{A}_{i+1}-1} \phi(\dot{V}_j^{(k)}, \dot{V}_{j+1}^{(k)})$ ,

we have for all  $n > 0$ ,

$$\frac{\sum_{i=0}^{\dot{N}_n-1} \dot{D}_i}{n} \leq \frac{\sum_{j=0}^n \phi(\dot{V}_j^{(k)}, \dot{V}_{j+1}^{(k)})}{n} \leq \frac{\sum_{i=0}^{\dot{N}_n} \dot{D}_i}{n}.$$

By setting  $\dot{C}_i = \dot{A}_{i+1} - \dot{A}_i$  for all  $i \geq 0$ , we have

$$\frac{\sum_{i=0}^{\dot{N}_n-1} \dot{C}_i}{\dot{N}_n} \leq \frac{n}{\dot{N}_n} \leq \frac{\sum_{i=0}^{\dot{N}_n} \dot{C}_i}{\dot{N}_n}.$$

The argument is essentially the same as for the proof of the *renewal reward theorem*. All the  $\dot{C}_i$  are independent and identically distributed (the Markov chain  $\dot{V}^{(k)}$  is homogeneous and  $\dot{V}_{\dot{A}_i}^{(k)} = \dot{q}$  for all  $i$ ). We put  $\mathbf{E}(\dot{C})$  for their expectation. Moreover, the chain  $\dot{V}^{(k)}$  is irreducible and contains a finite number of states, which are thus all positive recurrent. In particular, the mean recurrence time for  $\dot{q}$  is finite. Since, whatever  $i$ , the random variable  $\dot{C}_i$  accounts for the recurrence time of  $\dot{q}$ , the expectation  $\mathbf{E}(\dot{C})$  is finite, which implies that  $\lim_{n \rightarrow \infty} \dot{N}_n = \infty$ . The strong law of large number gives us that

$$\lim_{n \rightarrow \infty} \frac{n}{\dot{N}_n} = \mathbf{E}(\dot{C}) \quad \text{a.s.}$$

In the same way, the random variables  $\dot{D}_i$  are independent and identically distributed. Moreover, since  $\mathbf{E}(\dot{C})$  is finite and  $\phi$  is bounded, the identically distributed random variables  $\dot{D}_i$  have a finite expectation  $\mathbf{E}(\dot{D})$ . Applying again the strong law of large number leads to

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\sum_{i=0}^{\dot{N}_n} \dot{D}_i}{n} &= \lim_{n \rightarrow \infty} \frac{\sum_{i=0}^{\dot{N}_n} \dot{D}_i}{\dot{N}_n} \times \frac{\dot{N}_n}{n} \\ &= \frac{\mathbf{E}(\dot{D})}{\mathbf{E}(\dot{C})} \quad \text{a.s.} \end{aligned}$$

From the bounded convergence theorem, we get that

$$\begin{aligned}
\text{AS}_\pi^{(k)}(\Gamma_{\tilde{q} \circ \dot{q}}) &= \lim_{n \rightarrow \infty} \mathbf{E} \left( \frac{\sum_{j=0}^n \phi(\dot{V}_j^{(k)}, \dot{V}_{j+1}^{(k)})}{n} \right) \\
&= \lim_{n \rightarrow \infty} \mathbf{E} \left( \frac{\sum_{i=0}^{\dot{N}_n} \dot{D}_i}{n} \right) = \mathbf{E} \left( \lim_{n \rightarrow \infty} \frac{\sum_{i=0}^{\dot{N}_n} \dot{D}_i}{n} \right) \\
&= \frac{\mathbf{E}(\dot{D})}{\mathbf{E}(\dot{C})}.
\end{aligned}$$

The random variables  $C_{I_i^{\dot{q}}}$  are independent and identically distributed (with the same argument as above). Moreover, by construction, they follow the same distribution as the random variables  $\dot{C}_i$ . Since all the transitions that go to  $\tilde{q}$  in  $M$ , go to  $\dot{q}$  in  $\dot{M}$ , starting with  $\dot{q}$  and ending at the first  $\dot{q}$  or  $\tilde{q}$  in the chain  $V^{(k)}$  is the same as starting with  $\dot{q}$  and ending at the first  $\dot{q}$  in  $\dot{V}^{(k)}$ . The random variables  $C_{I_i^{\dot{q}}}$  have expectation  $\mathbf{E}(\dot{C})$ . In the same way, the random variables  $(D_{I_i^{\dot{q}}})_i$  are independent, identically distributed and follow the same distribution as the variables  $(\dot{D}_i)_i$ , thus have expectation  $\mathbf{E}(\dot{D})$ . The strong law of large numbers gives us

$$\begin{aligned}
\lim_{n \rightarrow \infty} \frac{\sum_{i=0}^{N_n^{\dot{q}}} D_{I_i^{\dot{q}}}}{N_n^{\dot{q}}} &= \mathbf{E}(\dot{D}), \quad \text{a.s.}, \\
\lim_{n \rightarrow \infty} \frac{\sum_{i=0}^{N_n^{\dot{q}}} C_{I_i^{\dot{q}}}}{N_n^{\dot{q}}} &= \mathbf{E}(\dot{C}), \quad \text{a.s.}
\end{aligned}$$

Moreover since the chain  $V^{(k)}$  is irreducible, we have

$$\lim_{n \rightarrow \infty} \frac{N_n^{\dot{q}}}{n} = \alpha_{\dot{q}}, \quad \text{a.s.},$$

where  $\alpha_{\dot{q}}$  is the probability of  $\dot{q}$  in the stationary distribution of  $V^{(k)}$ .

$$\begin{aligned}
\lim_{n \rightarrow \infty} \frac{\sum_{i=0}^{N_n^{\dot{q}}} D_{I_i^{\dot{q}}}}{n} &= \lim_{n \rightarrow \infty} \frac{\sum_{i=0}^{N_n^{\dot{q}}} D_{I_i^{\dot{q}}}}{N_n^{\dot{q}}} \times \frac{N_n^{\dot{q}}}{n} \\
&= \mathbf{E}(\dot{D})\alpha_{\dot{q}}, \quad \text{a.s.}
\end{aligned}$$

From the bounded convergence theorem, we get that

$$\begin{aligned}
\lim_{n \rightarrow \infty} \mathbf{E} \left( \frac{\sum_{i=0}^{N_n^{\dot{q}}} D_{I_i^{\dot{q}}}}{n} \right) &= \mathbf{E}(\dot{D})\alpha_{\dot{q}} \\
&= \frac{\mathbf{E}(\dot{D})}{\mathbf{E}(\dot{C})} \mathbf{E}(\dot{C})\alpha_{\dot{q}} \\
&= \text{AS}_\pi^{(k)}(\Gamma_{\tilde{q} \circ \dot{q}}) \mathbf{E}(\dot{C})\alpha_{\dot{q}}.
\end{aligned}$$

Symmetrically, we have

$$\lim_{n \rightarrow \infty} \mathbf{E} \left( \frac{\sum_{i=0}^{N_n^{\ddot{q}}} D_{I_i^{\ddot{q}}} }{n} \right) = \text{AS}_\pi^{(k)}(\Gamma_{\dot{q} \triangleright \ddot{q}}) \mathbf{E}(\ddot{C}) \alpha_{\ddot{q}}.$$

In order to prove that  $\mathbf{E}(\dot{C}) \alpha_{\dot{q}} + \mathbf{E}(\ddot{C}) \alpha_{\ddot{q}} = 1$ , let us define the random variables  $C_i = A_{i+1} - A_i$ . We have

$$\begin{aligned} \frac{\sum_{i=0}^{N_n} C_i}{n} &= \frac{\sum_{i=0}^{N_n^{\dot{q}}} C_{I_i^{\dot{q}}} }{n} + \frac{\sum_{i=0}^{N_n^{\ddot{q}}} C_{I_i^{\ddot{q}}} }{n} \\ &= \frac{\sum_{i=0}^{N_n^{\dot{q}}} C_{I_i^{\dot{q}}} }{N_n^{\dot{q}}} \times \frac{N_n^{\dot{q}}}{n} + \frac{\sum_{i=0}^{N_n^{\ddot{q}}} C_{I_i^{\ddot{q}}} }{N_n^{\ddot{q}}} \times \frac{N_n^{\ddot{q}}}{n}. \end{aligned}$$

Since the expectation of  $C_i$  is smaller than the expected return times of the positive recurrent states  $\dot{q}$  and  $\ddot{q}$ , it is finite. Since moreover

$$\frac{\sum_{i=0}^{N_n-1} C_i}{n} \leq n \leq \frac{\sum_{i=0}^{N_n} C_i}{n},$$

we have

$$\begin{aligned} &\lim_{n \rightarrow \infty} \mathbf{E} \left( \frac{\sum_{i=0}^{N_n} C_i}{n} \right) \\ &= 1 \\ &= \lim_{n \rightarrow \infty} \mathbf{E} \left( \frac{\sum_{i=0}^{N_n^{\dot{q}}} C_{I_i^{\dot{q}}} }{N_n^{\dot{q}}} \times \frac{N_n^{\dot{q}}}{n} \right) + \lim_{n \rightarrow \infty} \mathbf{E} \left( \frac{\sum_{i=0}^{N_n^{\ddot{q}}} C_{I_i^{\ddot{q}}} }{N_n^{\ddot{q}}} \times \frac{N_n^{\ddot{q}}}{n} \right) \\ &= \mathbf{E}(\dot{C}) \alpha_{\dot{q}} + \mathbf{E}(\ddot{C}) \alpha_{\ddot{q}}. \end{aligned}$$

The asymptotic speed of the recurrent class  $\mathcal{C}_k$  may be written as

$$\begin{aligned} \text{AS}_\pi^{(k)}(\Gamma) &= \lim_{n \rightarrow \infty} \mathbf{E} \left( \frac{\sum_{j=0}^n \phi(V_j^{(k)}, V_{j+1}^{(k)})}{n} \right) \\ &= \lim_{n \rightarrow \infty} \mathbf{E} \left( \frac{\sum_{i=0}^{N_n^{\dot{q}}} D_{I_i^{\dot{q}}} }{n} + \frac{\sum_{i=0}^{N_n^{\ddot{q}}} D_{I_i^{\ddot{q}}} }{n} \right) \\ &= \text{AS}_\pi^{(k)}(\Gamma_{\dot{q} \triangleright \dot{q}}) \mathbf{E}(\dot{C}) \alpha_{\dot{q}} + \text{AS}_\pi^{(k)}(\Gamma_{\dot{q} \triangleright \ddot{q}}) \mathbf{E}(\ddot{C}) \alpha_{\ddot{q}}. \end{aligned}$$

As a convex combination of  $\text{AS}_\pi^{(k)}(\Gamma_{\dot{q} \triangleright \dot{q}})$  and  $\text{AS}_\pi^{(k)}(\Gamma_{\dot{q} \triangleright \ddot{q}})$ ,  $\text{AS}_\pi^{(k)}(\Gamma)$  is smaller than their maximum. This last case leads again to

$$\text{AS}_\pi(\Gamma) \leq \max\{\text{AS}_\pi(\Gamma_{\dot{q} \triangleright \dot{q}}), \text{AS}_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}})\}$$

and ends the proof.  $\square$

**Corollary 1.** *Let  $\pi$  be an iid model,  $\Gamma$  be a standard  $w$ -matching machine. If the states  $\dot{q}$  and  $\ddot{q}$  are such that  $\mathbf{h}_\Gamma(\dot{q}) = \mathbf{h}_\Gamma(\ddot{q})$  then we have*

$$\text{AS}_\pi(\Gamma) \leq \max\{\text{AS}_\pi(\Gamma_{\ddot{q} \triangleright \dot{q}}), \text{AS}_\pi(\Gamma_{\dot{q} \triangleright \ddot{q}})\}$$

*Proof.* With Lemma 2 and Theorem 3, the sequences of states of an execution of  $\Gamma$  follows a Markov model which satisfies the assumptions of Theorem 6. From Lemma 1,  $\Gamma_{\ddot{q} \triangleright \dot{q}}$  and  $\Gamma_{\dot{q} \triangleright \ddot{q}}$  are still standard. Again with Theorem 3, the corresponding sequences of states follows two Markov models  $\dot{M}$  and  $\ddot{M}$ , respectively, which, by construction, satisfy the assumptions of Theorem 6.  $\square$

## 5.2 Minimal shift to a match - relevant states

Let  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  be a  $w$ -matching machine. A state  $q \in Q$  is *relevant* if it leads to a match transition reporting its current position, namely, if there exist a text  $t$  and two indexes  $i < j$  such that  $\mathbf{q}_\Gamma^t(i) = q$ ,  $\mathbf{q}_\Gamma^t(j) \in F$ ,  $t_{\mathbf{p}_\Gamma^t(j) + \alpha(\mathbf{q}_\Gamma^t(j))} = w_{\mathbf{q}_\Gamma^t(j)}$  and  $\mathbf{s}_\Gamma^t(k) = 0$  for all  $i \leq k < j$ . Under the implicit assumptions on matching machines (end of Section 3.1), all the pre-match states are relevant.

For all states  $q \in Q$ , we recursively define  $\text{minshift}(q)$  as:

$$\text{minshift}(q) = \begin{cases} 0 & \text{if } q \in F, \\ \min_{x \in \mathcal{A}} \{\text{minshift}(\delta(q, x)) + \gamma(q, x)\} & \text{otherwise.} \end{cases}$$

**Remark 6.** *If  $\Gamma$  is valid, then a state  $q$  is relevant if and only if  $\text{minshift}(q) = 0$ .*

Let  $\widehat{\Gamma}$  be the full memory expansion of  $\Gamma$  (Section 3.2). For all states  $q$  of  $\Gamma$ , we define  $\mathbf{a}(q)$  as the set comprising all the elements of  $R_{O_\Gamma}$  associated with  $q$  in  $\widehat{\Gamma}$ , namely,  $\mathbf{a}(q) = \{H \mid (q, H) \in \widehat{Q}\}$ . If  $\Gamma$  is standard then for all states  $q$ ,  $\mathbf{a}(q)$  is a singleton. A state  $q$  of  $\Gamma$  is said *consistent*, if all pairs  $(H, H')$  of elements of  $\mathbf{a}(q)$  verify the following properties

1.  $\mathbf{f}(H) = \mathbf{f}(H')$ ,
2. for all  $(i, x) \in H$ ,  $i \geq \text{minshift}(q) \Rightarrow (i, x) \in H'$ ,

where  $\mathbf{f}(H)$  is the set of position entries of the elements of  $H$  (see Section 3.2). In particular, all the states of a standard  $w$ -matching machine are consistent.

Two states  $q$  and  $q'$  are *interchangeable* if they are both consistent and such that all pairs  $(H, H')$  with  $H \in \mathbf{a}(q)$  and  $H' \in \mathbf{a}(q')$  verify the two properties just above.

**Lemma 6.** *Let  $\Gamma$  be a non-redundant  $w$ -matching machine in which all the states are consistent, and  $\dot{q}$  and  $\ddot{q}$  be two interchangeable states of  $\Gamma$ .*

1. *All the states of  $\Gamma_{\ddot{q} \triangleright \dot{q}}$  (resp. of  $\Gamma_{\dot{q} \triangleright \ddot{q}}$ ) are consistent.*
2. *If  $\Gamma$  is valid then both  $\Gamma_{\ddot{q} \triangleright \dot{q}}$  and  $\Gamma_{\dot{q} \triangleright \ddot{q}}$  are valid.*



3. If, moreover, for all states  $r$  and all symbols  $x$  and  $y$ ,  $\delta(r, x) = \delta(r, y) \Rightarrow \gamma(r, x) = \gamma(r, y)$  then, for all iid model  $\pi$ , we have

$$\text{AS}_\pi(\Gamma) \leq \max\{\text{AS}_\pi(\Gamma_{\dot{q}\triangleright\dot{q}}), \text{AS}_\pi(\Gamma_{\dot{q}\triangleright\ddot{q}})\}.$$

*Proof.* Property 1 comes straightforwardly with the definitions of consistency and interchangeability.

Property 2 may be proved in the same way as Lemma 1.

In order to prove Property 3, we remark that, since  $\Gamma$  is non-redundant and  $\dot{q}$  and  $\ddot{q}$  are interchangeable, both  $\Gamma_{\dot{q}\triangleright\dot{q}}$  and  $\Gamma_{\dot{q}\triangleright\ddot{q}}$  are non-redundant. With Theorem 3, we get that, if  $t$  is iid, the sequence of states parsed during an execution of the algorithm on the input  $(t, \Gamma)$  (resp.  $(t, \Gamma_{\dot{q}\triangleright\dot{q}})$ ,  $(t, \Gamma_{\dot{q}\triangleright\ddot{q}})$ ) follows a Markov model  $M = (\pi_M, \delta_M)$  (resp.  $\dot{M} = (\pi_{\dot{M}}, \delta_{\dot{M}})$ ,  $\ddot{M} = (\pi_{\ddot{M}}, \delta_{\ddot{M}})$ ). Moreover, since for all states  $r$  and  $s$ , we have

$$\delta_M(r, s) = \sum_{x, \delta(r, x)=s} \pi(x),$$

$$\delta_{\dot{M}}(r, s) = \sum_{x, \delta_{\dot{q}\triangleright\dot{q}}(r, x)=s} \pi(x),$$

$$\delta_{\ddot{M}}(r, s) = \sum_{x, \delta_{\dot{q}\triangleright\ddot{q}}(r, x)=s} \pi(x),$$

and with the definition of  $\Gamma_{\dot{q}\triangleright\dot{q}}$  and  $\Gamma_{\dot{q}\triangleright\ddot{q}}$ , both  $\dot{M}$  and  $\ddot{M}$  satisfy the assumptions of Theorem 6. If, moreover, for all states  $r$  and all symbols  $x$  and  $y$ ,  $\delta(r, x) = \delta(r, y) \Rightarrow \gamma(r, x) = \gamma(r, y)$ , all the assumptions of Theorem 6 are granted, which leads to Property 3.  $\square$

**Lemma 7.** *Let  $\Gamma$  be a valid, non-redundant and compact  $w$ -matching machine containing only consistent states. For all iid models  $\pi$ , there exists a  $w$ -matching machine  $\Gamma'$  such that*

1. *for all states  $q \in Q'$ , if  $\alpha'(q) < \text{minshift}(q)$  then for all symbols  $x$  and  $y$ , both  $\delta'(q, x) = \delta'(q, y)$  and  $\gamma'(q, x) = \gamma'(q, y)$ ;*
2.  $\text{AS}_\pi(\Gamma') \geq \text{AS}_\pi(\Gamma)$ .

*Proof.* Let us first remark that under the assumptions that  $\Gamma$  is valid and all its states consistent, if there exist two symbols  $x$  and  $y$  such that  $\delta(q, x) \neq \delta(q, y)$  then  $\gamma(q, x) \neq \gamma(q, y)$  (it can be proved in the same way as Lemma 2).

Let us assume that there exists a state  $q \in Q$  such that both  $\alpha(q) < \text{minshift}(q)$  and  $\delta(q, x) \neq \delta(q, y)$ . Since all the states are consistent, the states  $\dot{q} = \delta(q, x)$  and  $\ddot{q} = \delta(q, y)$  are interchangeable. Both  $\Gamma_{\dot{q}\triangleright\dot{q}}$  and  $\Gamma_{\dot{q}\triangleright\ddot{q}}$  are such that  $\delta_{\dot{q}\triangleright\dot{q}}(q, x) \neq \delta_{\dot{q}\triangleright\ddot{q}}(q, y)$ . Lemma 6 ensures that both  $\Gamma_{\dot{q}\triangleright\dot{q}}$  and  $\Gamma_{\dot{q}\triangleright\ddot{q}}$  contain only consistent states, are valid and such that

$$\text{AS}_\pi(\Gamma) \leq \max\{\text{AS}_\pi(\Gamma_{\dot{q}\triangleright\dot{q}}), \text{AS}_\pi(\Gamma_{\dot{q}\triangleright\ddot{q}})\}.$$

We put  $\Gamma'$  for the machine with the greatest asymptotic speed among  $\Gamma_{\tilde{q} \succ q}$  and  $\Gamma_{\tilde{q} \succ q}$ . If  $\Gamma'$  is such that for all states  $q \in Q'$ , if  $\alpha'(q) < \text{minshift}(q)$  then for all symbols  $x$  and  $y$ , both  $\delta'(q, x) = \delta'(q, y)$  and  $\gamma'(q, x) = \gamma'(q, y)$ , the lemma is proved. Otherwise, we replace  $\Gamma$  by  $\Gamma'$  which still satisfies the assumptions of the lemma and has a greater asymptotic speed before iterating the same process. Since at each iteration, there is a state  $q$  and two symbols  $x$  and  $y$  such that  $\delta(q, x) \neq \delta(q, y)$  and  $\delta'(q, x) = \delta'(q, y)$ , we eventually end with a machine  $\Gamma'$  with the desired property.  $\square$

Let  $\Gamma = (Q, o, F, \alpha, \delta, \gamma)$  be a  $w$ -matching machine verifying  $\alpha(q) \geq \text{minshift}(q)$  for all states  $q \in Q$ . The  $w$ -matching machine  $\Gamma^+ = (Q^+, o^+, F^+, \alpha^+, \delta^+, \gamma^+)$  is defined as

- $Q^+ = Q$ ,
- $o^+ = o$ ,
- $F^+ = F$ ,
- $\alpha^+(q) = \alpha(q) - \text{minshift}(q)$ ,
- $\delta^+(q, x) = \delta(q, x)$ ,
- $\gamma^+(q, x) = \gamma(q, x) - \text{minshift}(q) + \text{minshift}(\delta(q, x))$ .

for all states  $q \in Q$ .

If  $\Gamma$  is such that  $\alpha(q) \geq \text{minshift}(q)$  for all  $q \in Q$ , the quantities  $\alpha^+(q)$  and  $\gamma^+(q, x)$  are non-negative for all states  $q$  and all symbols  $x$  (i.e.  $\Gamma^+$  is well a  $w$ -matching machine).

**Remark 7.** For all texts  $t$ , the sequences of accessed positions coincide during the executions of the generic algorithm on the inputs  $(\Gamma, t)$  and  $(\Gamma^+, t)$ . In particular,  $\Gamma$  is valid if and only if  $\Gamma^+$  is valid and the asymptotic speeds of  $\Gamma$  and  $\Gamma^+$  are equal.

**Theorem 7.** Let  $\Gamma$  be a valid  $w$ -matching machine. For all iid models  $\pi$ , there exists a  $w$ -matching machine  $\Gamma_\pi$  with  $\text{AS}_\pi(\Gamma) \leq \text{AS}_\pi(\Gamma_\pi)$  and which is

- standard,
- compact,
- valid,
- in which all the states are relevant,
- such that there is no pair of states  $(q, q')$  such that  $q \neq q'$  and  $\mathbf{h}_{\Gamma_\pi}(q) = \mathbf{h}_{\Gamma_\pi}(q')$ .

*Proof.* With Proposition 2, there exists a standard, compact and valid  $w$ -matching machine  $\Gamma_a$  such that  $\text{AS}_\pi(\Gamma) \leq \text{AS}_\pi(\Gamma_a)$ . Since the machine  $\Gamma_a$  satisfies the assumptions of Lemma 7, there exists a  $w$ -matching machine  $\Gamma_b$  which is

- valid,
- in which all the states are consistent,
- with a asymptotic speed greater than  $\Gamma_a$ ,
- such that for all states  $q \in Q_b$ , if  $\alpha_b(q) < \text{minshift}(q)$  then for all symbols  $x$  and  $y$ , both  $\delta_b(q, x) = \delta_b(q, y)$  and  $\gamma_b(q, x) = \gamma_b(q, y)$ .

The machine  $\Gamma_b$  is still non-redundant. It is possibly non-compact but this can occur only in the case where there exists states  $q$  with  $\alpha_b(q) < \text{minshift}(q)$ . In this case, Lemma 3 may be applied, possibly several times, in order to get a compact and valid  $w$ -matching machine  $\Gamma_c$  with a greater asymptotic speed than  $\Gamma_b$ . Moreover,  $\Gamma_c$  does not contain any state  $q$  with  $\alpha_c(q) < \text{minshift}(q)$ .

Let us put  $\Gamma_d$  for  $(\Gamma_c)^+$ . All the states of  $\Gamma_d$  are relevant. If  $\Gamma_d$  is not standard and compact, Proposition 2 ensures that there exists a  $w$ -matching machine  $\Gamma_e$  which is valid, standard, compact, in which all the states are relevant and with a greater asymptotic speed than  $\Gamma_d$ .

Finally, applying Corollary 1 on  $\Gamma_e$  as long as there exist two states  $q \neq q'$  with  $\mathbf{h}_{\Gamma_e}(q) = \mathbf{h}_{\Gamma_e}(q')$ , eventually leads to a  $w$ -matching machine with the desired properties.  $\square$

**Corollary 2.** *Let  $w$  be a pattern,  $\pi$  be an iid model and  $n$  an integer greater than  $|w| - 1$ . Among all the valid  $w$ -matching machines of order  $n$ , there exists a machine  $\Gamma$  with a maximal asymptotic speed which verifies the properties of Theorem 7. In particular, it is standard, non-redundant and such that  $Q$  is in bijection with a subset of the partial functions  $f$  from  $\{0, \dots, n\}$  to  $\mathcal{A}$ , verifying that if  $f(i)$  is defined and  $i < |w|$  then  $f(i) = w_i$ .*

*Proof.* With Theorem 7, a valid  $w$ -matching machine of order  $n$  achieving the greatest asymptotic speed among the machines of order  $n$ , may be found among the  $w$ -matching machines  $\Gamma$  which are, among other properties,

1. standard,
2. such that there is no pair of states  $(q, q')$  such that  $q \neq q'$  and  $\mathbf{h}_{\Gamma_\pi}(q) = \mathbf{h}_{\Gamma_\pi}(q')$ ,
3. in which all the states are relevant.

The first property just ensures that the second one makes sense. The second property implies the bijection between the set of states and a subset of partial functions from  $\{0, \dots, n\}$  to  $\mathcal{A}$  by associating the states  $q$  with the partial function  $f_q$  corresponding to  $\mathbf{h}_\Gamma(q)$ . If for a state  $q$  and a position  $i < |w|$ , we have  $f(i) \neq w_i$ , then the state  $q$  is not relevant (or the  $\Gamma$  is not valid), which contradicts the property 3 (or the validity of  $\Gamma$ ).  $\square$

**Corollary 3.** *Let  $w$  be a pattern,  $\pi$  be an iid model and  $n$  an integer greater than  $|w| - 1$ . A  $w$ -matching machine achieving the greatest asymptotic speed among all the  $w$ -matching machines of order smaller than  $n$  can be computed in a finite time and with a finite amount of memory.*

	Naive	Morris-Pratt	Knuth-Morris-Pratt	Quicksearch	Horspool	FJS	TVSBS	EBOM	Hashq	Fastest
aaaa	0.753	0.803	1.000	1.705	2.324	1.393	1.255	1.385	0.651	<b>2.785</b>
aaab	0.753	0.823	0.996	0.536	1.480	0.581	0.306	1.134	0.633	<b>2.112</b>
aaba	0.736	0.839	0.985	0.747	0.810	0.739	1.038	0.914	0.625	<b>1.783</b>
aabb	0.736	0.856	0.973	0.627	0.475	0.606	0.352	0.706	0.578	<b>1.620</b>
abaa	0.674	0.815	0.921	0.901	1.214	0.952	1.156	0.914	0.627	<b>1.807</b>
abab	0.674	0.823	0.941	0.500	0.753	0.557	0.337	0.871	0.581	<b>1.560</b>
abba	0.634	0.823	0.901	0.756	0.885	0.719	0.604	0.610	0.556	<b>1.531</b>
abbb	0.634	0.874	0.874	0.613	0.486	0.654	0.411	0.432	0.461	<b>1.359</b>
baaa	0.504	0.575	0.575	1.001	1.788	1.059	0.980	1.134	0.631	<b>2.154</b>
baab	0.504	0.583	0.587	0.421	1.139	0.434	0.267	0.881	0.596	<b>1.504</b>
baba	0.481	0.583	0.640	0.524	0.810	0.646	0.792	0.871	0.575	<b>1.621</b>
babb	0.481	0.650	0.670	0.469	0.475	0.483	0.314	0.514	0.492	<b>1.154</b>
bbaa	0.408	0.635	0.655	0.733	1.214	0.851	0.771	0.706	0.567	<b>1.679</b>
bbab	0.408	0.665	0.703	0.358	0.753	0.418	0.303	0.514	0.485	<b>1.175</b>
bbba	0.366	0.698	0.760	0.450	1.032	0.812	0.533	0.432	0.407	<b>1.387</b>
bbbb	0.366	0.698	1.000	0.475	0.567	0.410	0.375	0.436	0.301	<b>1.241</b>

Table 1: Asymptotic speeds of standard algorithms for all the patterns of length 4 over  $\{a, b\}$  with  $\pi_a = 0.25$  and  $\pi_b = 0.75$ .

*Proof.* We first remark that, from Theorem 1, checking if a given  $w$ -matching machine is valid can be performed in a finite time. Computing its asymptotic speed with regard to an iid model  $\pi$  just needs to determine the limit frequencies of a finite Markov chain, which can also be done in a finite time.

Finally, since the subsets of partial functions from  $\{0, \dots, n\}$  to  $\mathcal{A}$  is finite, checking the validity and computing the asymptotic speeds with regard to an iid model  $\pi$  of all the  $w$ -matching machines of which the set of states is in bijection with a partial function from  $\{0, \dots, n\}$  to  $\mathcal{A}$ , can be performed in a finite time.  $\square$

Being given a pattern  $w$ , an iid model  $\pi$  and an order  $n$ , it is thus possible to determine with certainty a  $w$ -matching machine which achieves the greatest asymptotic speed, thus somehow the smallest asymptotic average complexity on texts following the distribution  $\pi$ . In the companion paper [?], we provide an algorithm for determining, being given any pattern  $w$ , an optimal  $w$ -matching machine of order  $|w| - 1$  with regard to a given iid model  $\pi$  (i.e. with the greatest asymptotic speed under  $\pi$ ). Table 1 displays the asymptotic speeds of some standard algorithms (see [4, 3]) and that of the optimal “fastest” one under a given iid model.

	Naive	Morris-Pratt	Knuth-Morris-Pratt	Quicksearch	Horspool	FJS	TVSBS	EBOM	Hashq	Fastest
GCAC	0.727	0.790	0.790	1.130	1.842	1.093	1.161	1.238	0.634	<b>2.377</b>
GACC	0.742	0.790	0.790	1.241	1.829	1.176	1.207	1.275	0.634	<b>2.405</b>
GTGT	0.744	0.789	0.824	1.297	2.057	1.190	1.193	1.304	0.637	<b>2.472</b>
GCAG	0.727	0.791	0.800	1.052	1.743	0.975	1.129	1.212	0.626	<b>2.094</b>
GACG	0.742	0.790	0.796	1.045	1.710	0.963	1.086	1.240	0.630	<b>2.105</b>
TTGC	0.758	0.823	0.945	1.144	1.892	1.008	1.037	1.199	0.639	<b>2.218</b>
CTTC	0.760	0.803	0.809	1.397	2.161	1.168	1.165	1.242	0.631	<b>2.321</b>
CGGC	0.743	0.805	0.813	1.348	1.990	1.112	1.110	1.164	0.623	<b>2.199</b>
CCAG	0.758	0.814	0.952	1.155	1.961	0.993	1.120	1.262	0.624	<b>2.232</b>
CTGC	0.751	0.805	0.819	1.096	1.691	0.978	1.084	1.181	0.621	<b>2.125</b>

Table 2: Average speeds of standard algorithms over *E. Coli* genome for 10 patterns randomly picked in the sequence.

## References

- [1] R. A. Baeza-Yates and M. Régnier. Average running time of the Boyer-Moore-Horspool algorithm. *Theoretical Computer Science*, 92(1):19 – 31, 1992.
- [2] G. Barth. An analytical comparison of two string searching algorithms. *Information Processing Letters*, 18(5):249 – 256, 1984.
- [3] C. Charras and T. Lecroq. *Handbook of Exact String Matching Algorithms*. King’s College Publications, 2004.
- [4] M. Crochemore and W. Rytter. *Text Algorithms*. Oxford University Press, 1994.
- [5] G. Didier and L. Tichit. Designing optimal- and fast-on-average pattern matching algorithms. <http://arxiv.org/abs/1604.08860>, 2016.
- [6] S. Faro and T. Lecroq. The Exact Online String Matching Problem: A Review of the Most Recent Results. *ACM Comput. Surv.*, 45(2):13:1–13:42, Mar. 2013.
- [7] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, 1968.
- [8] L. Guibas and A. Odlyzko. String overlaps, pattern matching, and non-transitive games. *Journal of Combinatorial Theory, Series A*, 30(2):183 – 208, 1981.

- [9] D. E. Knuth, J. H. Morris, Jr, and V. R. Pratt. Fast pattern matching in strings. *SIAM journal on computing*, 6(2):323–350, 1977.
- [10] H. M. Mahmoud, R. T. Smythe, and M. Régnier. Analysis of Boyer-Moore-Horspool string-matching heuristic. *Random Struct. Algorithms*, 10(1-2):169–186, 1997.
- [11] T. Marschall, I. Herms, H. Kaltenbach, and S. Rahmann. Probabilistic Arithmetic Automata and Their Applications. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 9(6):1737–1750, Nov. 2012.
- [12] T. Marschall and S. Rahmann. Probabilistic Arithmetic Automata and Their Application to Pattern Matching Statistics. In P. Ferragina and G. M. Landau, editors, *Combinatorial Pattern Matching*, volume 5029 of *Lecture Notes in Computer Science*, pages 95–106. Springer Berlin Heidelberg, 2008.
- [13] T. Marschall and S. Rahmann. Exact Analysis of Horspools and Sundays Pattern Matching Algorithms with Probabilistic Arithmetic Automata. In A.-H. Dediu, H. Fernau, and C. Martín-Vide, editors, *Language and Automata Theory and Applications*, volume 6031 of *Lecture Notes in Computer Science*, pages 439–450. Springer Berlin Heidelberg, 2010.
- [14] T. Marschall and S. Rahmann. An Algorithm to Compute the Character Access Count Distribution for Pattern Matching Algorithms. *Algorithms*, 4(4):285, 2011.
- [15] M. Régnier and W. Szpankowski. Complexity of Sequential Pattern Matching Algorithms. In M. Luby, J. D. Rolim, and M. Serna, editors, *Randomization and Approximation Techniques in Computer Science*, volume 1518 of *Lecture Notes in Computer Science*, pages 187–199. Springer Berlin Heidelberg, 1998.
- [16] R. T. Smythe. The Boyer-Moore-Horspool heuristic with Markovian input. *Random Struct. Algorithms*, 18(2):153–163, 2001.
- [17] T.-H. Tsai. Average Case Analysis of the Boyer-Moore Algorithm. *Random Struct. Algorithms*, 28(4):481–498, July 2006.
- [18] A. C.-C. Yao. The complexity of pattern matching for a random string. *SIAM Journal on Computing*, 8(3):368–387, 1979.